Aalto University
School of Business

LEVERAGING ASSOCIATION RULE MINING TO ACCELERATE SALES

Case Teknos Group Oy Paint Store Transaction Data

Master's Thesis
Noora Kuisma
Aalto University School of Business
Information and Service Management
Fall 2019

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Noora Kuisma

Title of thesis Leveraging association rule mining to accelerate sales

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Management

Thesis advisor(s) Merja Halme

Year of approval 2019 **Number of pages** 95 **Language** English

Abstract

Companies operating in challenging business environments, characterized by the proliferation of disruptive technologies and intensifying competition, are obliged to re-evaluate their strategic approach. This has become the norm in the retail industry and traditional brick-and-mortar stores. Particularly local market players with scarce resources are looking into alternative solutions to delivering a unique customer experience with the intention to preserve their profitability.

Customer experience has been an integral topic within academic research for decades, and has also substantiated its value in pragmatic contexts. Recent developments in this field have triggered the constitution of customer experience management functions, which aim to adopt a holistic approach to the customer experience. This enforces a quantitative perspective highlighting the role of customer transaction data.

Association analysis is one of the most well-known methodology used to detect underlying patterns hidden in large transaction data sets. It uses machine learning techniques to firstly identify frequently purchased product combinations and secondly, to discover concealed associations among the products. The association rules derived and evaluated during the process can potentially reveal implicit, yet interesting customer insight, which may translate into actionable implications. The practical consequences in the framework of this study are referred to as sales increasing strategies, namely targeted marketing, cross-selling and space management.

This thesis uses Python programming language in Anaconda's Jupyter Notebook environment to perform association analysis on customer transaction data provided by the case company. The Apriori algorithm is applied to constitute the frequent itemsets and generate association rules between these itemsets. The interestingness and actionability of the rules will be evaluated based on various scoring measures computed for each rule.

The outcomes of this study contribute to finding interesting customer insight and actionable recommendations for the case company to support their success in demanding market conditions. Furthermore, this research describes and discusses the relative success factors from the theoretical point of view and demonstrates the process of association rule mining when applied to customer transaction data.

Keywords Association analysis, Association rules, Machine learning, Transaction data, Customer experience, Customer experience management, Customer insight, Targeted marketing, Cross-selling, Space management, Data mining

Acknowledgements

I would like to express my gratitude to Teknos for providing me with the opportunity to conduct this research on a topic that not only was particularly interesting for me, but also, benefited my academic studies. A special thank you goes to Marcel Dissel for always encouraging innovative thinking, trusting my judgement, and supporting me throughout the research process. In addition, I wish to thank Peter Svensson for providing me with real customer data, an invaluable asset to this thesis. I great thank you also for sharing extensive product and business knowledge, without which the results and analysis of this research would have fallen short.

Above all, I want to thank my thesis supervisor, Merja Halme, who has offered her guidance and advice during the most critical moments. Her dedication, passion, and perseverance have had a great impact on the whole process of this study, and have served as the primary sources of motivation.

Finally, I thank the Aalto University School of Business for the excellent education that has contributed to the overall extraordinary experience of study, and my friends and family who have supported me during this journey. Without their unconditional faith and enthusiasm, these past years would have been challenging to say the least.

Table of Contents

List of Tables

List of Figures

1 INTRODUCTION

Recent obstacles hindering retail network development and the success of traditional brick-and-mortar stores have contributed to the indispensable urgency to re-evaluate the business models of companies operating in this environment. Taking a holistic approach to customer experience by combining both qualitative and quantitative information on customer purchase behaviour has become one of the prerequisites for successfully overcoming these challenges. This encourages simultaneously improving customer satisfaction and taking action to increase in-store sales. (Reinartz & Kumar 1999; Verhoef et al. 2009; Brodie et al. 2011; Agnihotri 2015; Lemon & Verhoef 2016).

Albeit current developments in academic literature and corporate environments indicate apparent course of action, a comprehensive transfer from grasping the customer experience to strategically managing the customer experience is not a straightforward undertaking. Not only does this require major organizational restructuring, but also, significant effort to integrate customer data and analytical capabilities. This consolidation plays a fundamental role in the process of transferring customer information into customer knowledge and ultimately, the customer knowledge into actionable customer insight. (Homburg et al. 2015; Lemon & Verhoef 2016; Al-Rubaiee et al. 2018).

The objective of this thesis is to support this process by utilizing quantitative customer transaction data and performing association analysis. Hence, two significant outcomes emerge. Firstly, important customer insight is gained. Secondly, targeted marketing, cross-selling and space management solutions can be derived. These advancements in principal contribute to improving customer satisfaction and stimulating sales growth. Furthermore, this research supports the structural re-organization of the case company and the mitigation of the threats associated with the business landscape.

1.1 Background

This section will shed light on the current global economic trend influencing the survival of brick-and-mortar (B&M) stores. This part has an essential role in the interpretation of the motivation of this research, and thus, the formulation of the research questions. Furthermore, the environment described here acts as the foundation for the literature review, which construes the survival strategies taken in response to this trend.

The retail industry and the B&M stores have existed amidst many fluctuating developments, one of the most disruptive being internet shopping and multichannel approaches (Reinartz & Kumar 1999; Agnihotri 2015). A recent study in the United States indicates an astounding annual growth rate of 20% within online retail, whereas B&M retail has grown at moderate 4% rate during the time period of twelve years (Morgan Stanley 2013). This suggests that competitive pressure from online stores has been the dominating element contributing to the eroding growth prospects of traditional B&M stores.

The prevailing disruptive trends challenging the traditional B&M stores are discernible in the architectural coatings industry as well. Companies operating in the industry are committing to conflicting survival strategies, depending primarily on the organizational resources available. Multinational market leaders, such as Akzo Nobel and PPG, are constantly pursuing new innovative strategies aimed at revamping the in-store experience of the end consumer. These tactics range from developing digital color design tools to opening unconventional collaborative showrooms. The most revolutionary approaches involve recreating the entire business model by adopting multichannel strategies, which emancipate the sale of coatings from the physical store premises to an online environment. (“Different perspectives” 2019; “New PPG shipping hub” 2019).

However, smaller national companies with sparse resources and limited brand awareness are often obliged to withdraw from the market. For instance, Tikkurila Group Oyj, a Finnish coatings manufacturer and supplier, recently announced closing yet again another fraction of their paint store network as part of their cost savings initiatives (“Tikkurila myy” 2019). Also other local paint suppliers have been obliged to sell their physical store divisions to

larger global competitors so as to reduce operational costs in demanding market conditions ("PPG acquires paint stores network" 2019).

1.2 Motivation

Given a two-tier competitive landscape with a selected few global players and a number smaller local manufacturers, the latter face a tough decision in the occurrence of disruptive technologies. They can choose to enforce a defensive strategy, and sell the B&M stores to their global counterparts with adequate resources and competences to take the stores to the next level, while remaining at the top of the local industry. The other alternative is to employ an offensive strategy, and utilize the existing resources in a totally unique way.

The motivation for this research is to induce an offensive approach by providing the case company with insight to enable tackling the challenges associated with the B&M stores. These survival strategies are twofold: firstly, this research aims to identify interesting customer insight, that can be used to improve the in-store customer experience. Secondly, the topic of this research is motivated by finding solutions that are specifically targeted at increasing sales of the B&M stores. However, these solutions can also be utilized to increase sales in online environments, if the case company decides to expand their business into alternative channels.

Partly resulting from the vital need to discover survival tactics in the challenging market situation and partly due to the strategic redesign of the case company, another motivation to study this topic emerges: to capitalize on existing unique resources in an unprecedented manner. In essence, the customer transaction data from the paint stores of the company has never been extracted, analyzed and leveraged. Therefore, this research also addresses the need to exploit the existing company resources in novel ways, without having to commit to extensive investments.

Moreover, the motivation behind this thesis derives from an identified gap in academic literature. A lot of earlier academic research on association analysis techniques exist especially in the fields of computer science and medical research. These methods have recently been applied to customer transaction data as well, but are mainly limited to the retail

industry, and especially grocery stores in the United States. No earlier research in the literature of association analysis, customer transaction data analysis, or market basket analysis can be found for paint stores.

1.3 Research questions

Based upon the predominating economic background and the motivations of this research, the research questions can be defined. These are as follows:

Research question 1: Which product combinations are frequent?

This question aims to identify the most commonly purchased item combinations. Identifying the most frequent product combinations is a prerequisite for the second research question, which examines the associations between the products. Despite this, this research question provides also added value for the case company as it reveals information that is previously unknown. To answer this question, the most frequent item combinations, or itemsets, will be extracted from the customer transaction data.

Research question 2: What kind of associations can be detected between different products and product combinations?

After answering the preceding research question, the associations can be detected. This will be accomplished by a process of association rule generation. In addition to the type of the discovered associations, i.e. the direction of the rules, also the quality of the associations will be assessed. With regards to the quality of the associations, various scoring methods will be assigned, including both objective and subjective measures.

1.4 Thesis structure

The first part of this thesis has introduced the research topic and the background conditions that have impacted the motives to conduct this study. The research questions were also presented. The following chapter will construct the theoretical part of this thesis, followed

by an introduction to the case company. Next, the methods will be introduced, namely the academic literature on association analysis and its application to this research.

Succeeding the theoretical background and terminology related to association analysis, the data will be described. This section commences the empirical part of this study, and includes an explanation of the cleaning and preparation process of the customer transaction data. Following is the association analysis and the obtained results. The Discussion chapter concludes the empirical part of this thesis, including revisiting the research questions, proposing managerial implications and recommending topics for future research.

2 THEORETICAL BACKGROUND

In this part I elaborate on previous academic literature critical to the interpretation of the objectives of this research and the case company's strategical reorganisation. The theoretical background also plays an integral role in selecting the appropriate methodology for the study and discussing the managerial implications that follow. Consequently, the first part of this literature review will examine the definitions of customer experience, an inaugural starting ground for comprehending the establishment of customer experience management. Next, customer transaction data and its function in turning customer insight into actionable sales increasing strategies will be reviewed.

The second part of this theoretical foundation will focus on earlier research on sales increasing strategies, principally targeted marketing, cross-selling and space management. These strategies are particularly applicable in the context of association analysis, which will be explained in the section on Methods. Although the theoretical background of this study is composed of two larger entities, it is important to acknowledge the causal relationship here. In other words, developing a supreme customer experience also impacts company sales. Hence, customer experience development and management can be considered a prerequisite for increasing sales, and therefore it is sensible to begin with outlining customer experience.

2.1 Developing customer experience

Despite the overwhelming quantity of evidence pointing towards the extinction of B&M businesses, their supremacy in contrast to competing multichannel approaches remains unquestionable in one respect. This is the ability to create a unique shopping experience for the customer, which has recently also gained attention in pragmatic settings. For instance, KPMG, Google and Amazon report having appointed customer experience officers, customer experience managers and other functions responsible for the strategic customer experience management ("Improving customer experience" 2019). Correspondingly, it is claimed that the customer experience management initiatives have driven the success of Starbucks. This success has followed the distinctive strategies and practices employed,

facilitating Starbucks to differentiate from competition. (Enders & Jelassi 2000; Michelli 2009; Verhoef et al. 2009; Mehra et al. 2013; Agnihotri 2015).

2.1.1 Customer experience – what is it?

Schmitt (1999) defines customer experience as a combination of five different types of experiences: sensory, affective, cognitive, physical and social-identity. Verhoef et al. (2009) also take a multidimensional approach and conclude that the customer experience concept is holistic in nature and includes the customer's cognitive, affective, emotional, social and physical reactions to the retail company. These definitions align with various theories and descriptions related to brand experience. For example, Schmitt et al. (2014) propose four different components that form the brand experience: the sensory, affective, and the intellectual.

Later on, the scope of the various customer responses, or clues, considered in the whole customer journey was broadened, and a relationship approach was adopted. These developments introduced a data driven angle to the customer experience research. This customer centric methodology, known as customer relationship management (CRM), aims to collect, analyse and utilize customer data from all touchpoints in the customer journey and to embed this data deep into the organizational functions and operations. (Abbott et al. 2001; Gebert et al. 2003; Berry et al. 2006; Shihab et al. 2015; Homburg et al. 2015; Lemon & Verhoef 2016).

It is argued that the CRM construct acted as the antecedent to the development of customer experience management (CEM) (Homburg et al. 2015; Lemon & Verhoef 2016; Al-Rubaiee et al. 2018). CRM and CEM have many comparable characteristics since both employ a strategic approach to managing the customer experience through data usage, creating value to both the company and the customer. However, some salient dissimilarities emerge. Lemon and Verhoef (2016) define CEM as a value-driven strategy framing the experience of the customer by leveraging data of the current customer experience, whereas CRM focuses on recording data from all historical customer experiences. Furthermore, it is argued that CRM concentrates on value extraction, while CEM emphasizes value creation (Lemon & Verhoef 2016).

2.1.2 Customer experience management

The previous academic research explicitly focusing on customer experience management is rather finite, and in principal, concentrates on authoritative actions and implications (Berry et al. 2006; Verhoef et al. 2009; Lemon & Verhoef 2016). Schmitt (2018) defines CEM as "the process of strategically managing a customers' entire experience with a product or company". Homburg et al. (2015) define CEM as "the cultural mindsets toward customer experiences, strategic directions for designing customer experiences, and firm capabilities for continually renewing customer experiences, with the goals of achieving and sustaining long-term customer loyalty". CEM was initially considered as an integral managerial topic with only moderate focus on its relevance to the organization as a whole.

The strategic emphasis of CEM highlights big data and analytical capabilities (Wedel & Kannan 2016). These competences have key responsibility when aiming to firstly understand the customer journey, and secondly, to succeed in providing personalized service or product offerings for the customer. Thus, developing a structured method that systematically collects, captures, and analyses customer data from each stage of the customer journey can provide the company with important insight into the purchase habits of customers. The role of customer transaction data mining in supporting CEM as well as encouraging sales growth will be discussed in the next section.

2.1.3 Customer transaction data mining

Recent studies indicate an increasing amount of focus shifting towards the retail industry, their in-store operations, and the data being engendered there (Sands et al. 2009; Antczak & Weron 2019). The terms transaction data, point-of-sales (POS) data, point-of-purchase data and sales receipt data are frequently used to describe the data generated during the checkout operation of a customer, see for example Zhu (2013), Antczak & Weron (2019) and Tolbert (2008).

Data mining can be used to gain valuable insight from customer transactions to enable incorporating the information into various sales increasing strategies. Chen et al. (2006) summarize the concept of data mining as a process that enables detection of anteriorly

unknown patterns from a large data set by selecting, exploring, and modelling practices. Divulging obscure patterns for instance in customer behaviour can provide companies with a compelling competitive advantage. Moreover, mining customer transaction data is a key constituent when aiming to place the customer into the centre of all company operations, a focal mission of CEM (Tolbert 2008; Hsu et al. 2012; Wedel & Kannan 2016).

2.2 Increasing sales

While it is vital to appreciate the notion of customer experience and how it has evolved into CEM, the cardinal objective of this research is to place emphasis on the practical strategies used to drive sales. In this study, the following three sales increasing strategies are explored: targeted marketing, cross-selling and space management. These tactics exploit interdependencies between products and product combinations and therefore are especially applicable in the context of association analysis. Moreover, these approaches are focused on due to their relatively low initial investment requirements and their applicability to the case company conditions.

2.2.1 Targeted marketing

Targeted marketing actions refer to proposing the most convenient product or service to a customer by taking into consideration the unique characteristics, behavioural patterns and needs of the customer (Hwang & Yang 2008; Hardoon & Shmueli 2013). The behavioural patterns of the customer, as well as the underlying needs and wants, can be revealed by analysing the purchase preferences of the customer. These buying habits in turn can be unveiled by conducting qualitative customer surveys or by quantitative customer transaction data mining.

Targeted marketing initiatives in practice may refer to newsletters, product recommendations or targeted campaigns. Companies have long recognized the value of targeted marketing from the profitability perspective, as it has great potential in reducing marketing costs (Hwang & Yang 2008; Hwangbo et al. 2018). Another unquestionable contributing factor reducing the marginal costs of marketing, is the proliferation of available

customer data and use of analytical tools (Cooil et al. 2008; Tianyi & Tuzhilin 2009). This is particularly evident in ecommerce businesses and the use of social media as an advertisement channel. The additional cost of sending a newsletter to one more customer, say via email, is close to null, whereas the circumstances in more traditional marketing channels is very different (Cooil et al. 2008; Hwang & Yang 2008; Tianyi & Tuzhilin 2009).

Despite the decrease of marginal marketing costs, and hence the potential decrease of emphasis on the cost side of marketing, empirical evidence shows that targeted marketing actions may improve profitability also through other means. Firstly, customers are often more responsive to promotions if their prior preference for the promoted product or product group is higher. Thus, promoting a product or brand that the customer has shown to prefer in the past, suggests that the customer has a positive response to a future marketing campaign targeted at that specific or similar product. Secondly, behavioural studies have shown a link between personalized product offerings and customer satisfaction. These studies show that customers enjoy getting individualized offerings, insinuating that the customer is considered as an important, unique entity. (Khan et al. 2009; Reutterer et al. 2016).

Rossi et al. (1996) have investigated the profitability impacts of targeted marketing actions in light of historical purchase data and targeted couponing. Their findings suggest that the prospective benefits of utilizing customer data on past buying behaviour entail a net increase in revenue 2.5 times that of a “blanket strategy”, referring to mass-marketing of coupons. This gain was observed even with a confined set of past purchases, averaging only 13.23. In fact, it was revealed that even when considering only one historical purchase record to identify the optimal target group, net couponing revenue was improved by 50% compared to alternate tactics. (Rossi et al. 1996).

Targeted marketing efforts are perhaps most prevalent in the context of recommendation algorithms used in e-commerce. For instance, Amazon uses item-to-item collaborative filtering as a marketing tool to create a personalized shopping experience for their customers. Item-to-item collaborative filtering computes the associations between previously purchased products and other similar items to compose a list of recommendations for the customer. This filtering technique differs from traditional clustering and search-based methods in the sense that it exploits the connections between purchased items and other resembling products instead of matching the customers against each other. (Linden et al. 2003).

Although the emphasis here has been in the online store environments, targeted marketing initiatives are applicable also in offline store premises. The B&M stores can seize these opportunities by providing personalized customer service or designing personalized shopping recommendations. Also, these strategies can be scaled to service portfolio design and customized service recommendations. (Rossi et al. 1996; Khan et al. 2009; Linden et al. 2003)

Given these circumstances, it is beneficial to focus on positively correlated product pairs. Thereby, the company can enable personalised offerings by promoting a product with strong association to customers who have previously purchased the associate product. This reduces the marginal marketing costs, since the number of customers to whom the offer is targeted is bounded. The return on investment (ROI) also improves, given the increase of likelihood of the customer responding to the promotion. Even more so, the customer satisfaction increases, and therefore, also the in-store sales.

2.2.2 Cross-selling

Targeted marketing has important implications also for cross-selling purposes. These two theorems are similar in the sense that both recognize the unique needs and wants of the customer, are often consolidated with customer segmentation practices, and are closely associated with a company's marketing functions. Cross-selling tactics however, are often aimed at existing customers, whereas targeted marketing actions can be applied to both new and existing customers, depending on the initial contact channel.

From a company's perspective, cross-selling can be defined as a strategical approach to offering and selling supplementary services or products to an already existing customer. For example, a company with an extensive customer base may want to promote usage of their entire product or service offering so as to capitalize on the existing customer relations. Alternatively, a company may want to identify additional product features valuable to the customer in order to extend these characteristics to other similar products in the product portfolio, and to emphasize these qualities in their marketing practices. (Knott et al. 2002; Li et al. 2011; Malms & Schmitz 2011).

From a customer's perspective, cross-selling can be thought of either as a personalized offer that is custom designed to fit the customer's needs, or the opportunity to purchase services or products from the full assortment of the company. The guiding principle behind cross-selling is therefore to offer a wider assortment of offerings, increase sales, and engage the customer (Li et al. 2011; Malms & Schmitz 2011).

The practical implications for successfully performing cross-selling initiatives thus require that the sales representatives in the company have a wide knowledge of the variety of services or products sold. In addition, collaboration and knowledge sharing within the organization in considered a prerequisite for cross-selling. It is also imperative that the company knows their customers in detail. As straightforward and unambiguous as these conditions might seem at first glance, they are actually the principle causes of the failures of cross-selling projects (Homburg et al. 2019). Based on this discussion, the fundamental challenges can be divided into roughly two: the inter-organizational challenges related to the culture, motivation and attitudes within the company, and the obstacles related to the customer knowledge.

The inter-organizational issues refer to vigorous silo mentalities, especially prevailing in larger organizations. These attitudes may lead to hesitancy to sharing product or customer knowledge in fear of losing ownership of certain product lines or customers (Li et al. 2011; Homburg et al. 2019). This is an extremely dangerous pitfall, as it may lead to a situation where organizational silos are created even into structures where they didn't exist prior to the cross-selling process. Previous academic literature has approached the challenges of cross-selling particularly from this viewpoint, and improvement suggestions to advance cross-selling often relate to these inter-organizational issues.

Customer knowledge related challenges often arise when the company lacks the relevant customer knowledge or uses the wrong approach to target the customers. Li et al. (2011) allege that many cross-selling projects are designed and implemented with the goal of finding and targeting those customers who are most likely to respond to the campaign. In these scenarios, the companies first set a time constraint and then decide on a communication channel. Next, the customer responses to the campaign are estimated based on the product ownership and data of the customer. Finally, the expected profit is computed, and the promotions are targeted accordingly, given potential budget constraints.

Instead of aiming for an aggressive cross-selling approach designed from the company's point of view, organizations most likely benefit more by taking customer-centric orientation (Li et al. 2011). This customer-driven strategy resonates with the objectives expressed relating to the holistic customer experience management principle as well. Li et al. (2011) add that companies aiming to leverage cross-selling strategies should primarily aim to answer the following question: "how do we introduce the right product to the right customer at the right time using the right channel to ensure long term success?". This contrary view takes a step forward from the company-centric mentality targeting at the maximal response rate of the cross-selling project.

Knott et al. (2002) recognize the negative impact of poorly executed cross-selling initiatives and examine different customer-centric methodologies intended at improving these inefficiencies. In their research on cross-selling models within the banking industry, they consider three types of customer-specific information, including the demographic data, the monetary value and the product ownership of the customer. Product ownership here refers to the number of accounts that the customer "owns" in each product category, i.e. describes the previous purchase records of the customer. The results of the study indicate that all three attributes enhance the efficiency of cross-selling tactics, the purchase preferences of the customer being unequivocally the most important contributing factor. These findings align with the conclusions drawn from Rossi et al. (1996) study on targeted couponing.

2.2.3 Space management

The final sales increasing strategy is store space management, and particularly the product placement decisions involved. This is the third element discussed in the results of this research aiming to increase the sales. Space management can be defined as the optimal allocation of space to a certain product or product group to achieve a specific goal (Kollat & Willet 1967; Abratt & Goodey 1990). Space management techniques comprise of several schemes, a majority of the attention being assigned to decisions concerning product placement optimization.

Managing the in-store space has been utilized in optimization problems for long. Product placement issues are predominantly studied for instance in warehouse management and

shelf-space allocation settings. However, it has only recently gained attention in the retail industry, and especially in the in-store block layout design. (Ozgormus & Smith 2018).

The goals for managing store space vary depending on the company's intention, but are often focused on either improving the operational efficiency of the store or increasing impulse purchases (Kollat & Willet 1967; Abratt & Goodey 1990; Ozgormus & Smith 2018). The operational efficiency of the store layout can be viewed from both the customer's perspective and the company's perspective. An operationally efficient store from a customer's perspective may be one that is convenient to navigate in, and perhaps one that minimizes the length of the path needed to travel between the entrance and the exit. From the company's standpoint the store efficiency could be improved by placing fast mover items, or those that are purchased frequently, close to the warehouse, in order to minimize travel distances between the store and warehouse.

In addition to improving operational efficiency of a store by manipulating the product space allotment or store layout design, space management strategies place considerable focus on encouraging impulse purchases (Kollat & Willet 1967; Abratt & Goodey 1990). According to Kollat and Willet (1967), impulse purchases account for an astounding 30% to 50% of all customer purchases. Furthermore, researchers assert that it is specifically the store layout and design, including various visual elements, that affect these impulse purchases (Ozgormus & Smith 2018).

Certain strategies seizing store space design to encourage impulse purchases are rather obvious. Consider for instance furniture giant IKEA. The in-store space configuration aims to guide the customer through each individual room, showcasing the carefully planned interior design, without explicitly offering a shortcut to decrease the in-store travel distance. The underlying logic relies on the assumption that the likelihood of impulse purchases increases when being exposed to visually pleasing elements and a larger variety of products (Kollat & Willet 1967; Abratt & Goodey 1990).

A similar logic is used in the proximity of check outs in various retail stores. Here the customer is directed through racks and shelves displaying everyday necessities that are easy to grab while waiting in line. This tactic aims to take advantage of the excess time at the customer's hands, and the often resulting awareness of lacking a certain necessity. However,

exploiting this space management plot might lead to controversial results when the customers' preferences are not truly apprehended. In fact, some studies indicate deterioration in the overall customer experience in cases where a store aims to maximize the number of items presented to the customer by forcing the customer to navigate through the entire store. This in turn decreases the likelihood of supporting the company sales by engaging in impulse buying, as the overall experience has declined. (Hui et al. 2013; Ohmori et al. 2019).

3 CASE COMPANY

This chapter will briefly introduce the case company and the industry it operates in. Next, the research problems will be introduced from the case company perspective by reviewing the company's journey towards customer experience management and its relative success with regards to the challenges related to the B&M stores.

The case company for which this research is conducted, Teknos Group Oy, is a family-owned coatings manufacturer from Finland, offering a wide range of products for the manufacturing industry, building professionals and end consumers. The company employs roughly 1700 people in over 20 countries worldwide, and had a turnover of 408 million euros in 2018. Teknos states their mission as follows: "to make the world last longer by providing smart, technically advanced paint and coatings solutions to protect and prolong" ("We make the world last longer" 2019).

3.1 Paint and coatings industry

The paint and coatings industry is a traditional, mature industry with moderate growth prospects of 2% to 5% per year. The market growth in areas recently opened to Western commercialization, for instance Eastern Europe and the Asia Pacific region, is forecasted to develop more rapidly. Especially countries like China and India are heavily investing in their infrastructure development, and as a result, also the global paint and coatings manufacturers are gaining foothold there. (Weiss 1997).

The number of coatings manufacturers has significantly declined (approximately 30%) during the past years, and is projected to continue decreasing in the future as well. This is a common trend in mature industries, and is a consequence of the vast number of merges and acquisitions of key global manufacturers. A two-tier market has emerged, consisting of a selected group of global players, and local manufacturing companies targeting local niche markets. (Weiss 1997).

Weiss (1997) summarizes the division of the paint and coatings market into three broad segments: architectural or decorative coatings, industrial coatings and specialty coatings. The architectural coatings segment consist of paints, varnishes and lacquers that are applied directly to the surface, whether it be interior or exterior. These paints typically include house paints, stains and undercoats. The industrial coatings are applied during a specific, predetermined manufacturing process, most commonly within the facilities of the manufacturing company. These durable coatings are specifically designed according to the customer's specifications, and can be further grouped based on the material they are applied on i.e. industrial wood or metal. The specialty coatings are formulated to withstand extreme conditions, such as high abrasion, corrosion or temperature. (Weiss 1997).

3.2 Reducing silos with customer experience management

Historically, the operations of Teknos firmly correlated with the general paint and coatings market segmentation. They divided their products into four distinct segments: architectural coatings, powder coatings, industrial wood and general industry. This classification was primarily based on the technical properties of the product and the surface material of the area of application. However, recently it has been acknowledged that this technology-centric view does not consider the customer, or put their needs in the focal point of the company's activities.

Consider for example a customer manufacturing furniture. One specific piece of furniture is often made of various different materials i.e. wood, plastic and metal. This scenario leads to multiple challenges. First, the customer does not necessarily know the technical details of the manufacturing material of the furniture, and therefore is unaware of which type of coating is best suited for each part. Second, having detailed knowledge of the different surface materials does not imply that that the customer initially knows which coatings manufacturer to approach for each type of coating. Furthermore, the worst case scenario leads to a situation where the customer has to interact with multiple suppliers, complicating and hindering the manufacture process. A strict segmentation approach that assigns the customers into the previously mentioned four categories potentially also reduces cross-selling opportunities, and deepens organizational silos. Therefore, the customer experience management function was created.

Teknos updated their strategical initiatives for years 2019-2025 to support their ambitious goal of increasing their revenue to 1 billion euros. One of the strategic themes for the next six years was to improve the customer experience. As part of this strategic theme, Teknos quickly adopted the idea of a new organisational culture and ways of working. It was realized that the traditional narrow-minded segmentation does not address the dynamic nature of today's business world and the increasing customer demands. It certainly did not fit the vision of placing the customer in the centre of all operations, which is why the new Customer Experience Management (CEM) function was created, and new Target Customer Groups (TCG) were put in place accordingly.

The aim of creating the CEM organizational function and the new TCGs is to optimize all interactions between Teknos and their customers in a manner that most benefits the customer. Instead of obliging the customer to adjust to Teknos' segments, Teknos will create a strategy that conforms to the customer's operations. By doing this, Teknos is able to offer the customer an expansive product portfolio all from one supplier.

3.3 Challenges in B&M stores

Teknos has developed its B&M store network in various global markets, the majority of it residing in Sweden. Their store network growth is primarily a consequence of acquisitions, but is recently also impacted by opening entirely new stores under the brand name. Particularly the latter have been affected by the deteriorating global market, especially in locations where the brand perception is limited.

The paint stores of Teknos cater generally to the end consumer, but are not confined to this customer segment. The end consumer here may refer to home owners, home renovators, renovating companies and professional painters. In the case of the store network in Sweden, the majority of the customers belong to the professional painter category. This category can further be subdivided into small, medium and large painting companies.

The data used in this thesis has been collected from the oldest store located in Linköping, Sweden. The store was established in 2008 to serve the local market of professional painters. During the past 11 years the store and brand name has succeeded in gaining foothold in

competition, but this has not always been the case. Initially the store struggled to prove its value in an area of low brand awareness, a trend that continues to manifest itself in the retail industry.

4 METHODS

This chapter describes the literature and terminology of the methods used in this thesis. Firstly, association analysis as a process will be defined. Next, frequent itemset generation and the algorithm applied here will be illustrated. Alternative approaches and algorithms will also be introduced. Following the frequent itemset generation, association rule generation and evaluation will be discussed. The evaluation metrics will be described each in detail, as they compose a significant element in the results analysis.

4.1 Association analysis

Given a large set of customer transactions, it is enviable to discover different associations between products and product groups (Chen et al. 1996). Association analysis uses machine learning techniques that employ a variety of mathematical algorithms to reveal these implicit affiliations (Zaki 2000; Zheng et al. 2001; Larose & Larose 2005; Tan et al. 2005). The applications of association analysis within the field of business include process re-engineering, fraud detection, market basket analysis and item placement, the latter two being relevant to this study.

More often than not, three severe concerns arise when applying association analysis to customer transaction data:

1. Firstly, the number of different combinations of items in customer transaction data instantaneously add up when growing the number of overall items sold (see equation 1)
2. Secondly, increasing the number of items included in a specific itemset proliferates the amount of association rules generated (see equations 2 and 3). This often requires a lot of computational resources, and can be extremely time consuming. Having even the required computational power to perform such in-depth analysis does not guarantee successful results. This leads to the third major issue in applying association analysis to transaction data.
3. The discovered patterns between the different itemsets may either occur by chance, or otherwise be completely irrelevant or uninteresting. (Tan et al. 2005).

To address these issues, association analysis techniques are often subdivided into three smaller problems. These subtasks begin from the generation of frequent itemsets, then proceed to constructing the association rules, and finally, employ techniques to evaluate and prune the produced rules (Zaki 2000; Zheng et al. 2001; Larose & Larose 2005; Tan et al. 2005). The following sections will take a closer look at each step and provide numerical examples in the context of customer transaction data for illustrative purposes.

4.2 Frequent itemset generation

The number of different product combinations that customers can potentially purchase are immense, even with a narrow product portfolio. Let's demonstrate this with a numerical example. Consider a small store selling only 20 different unique items. Assuming that the *transaction width*, or the number of items present in a transaction is five, the different itemset combinations overall therefore add up to:

$$\frac{20!}{(5!\,(20-5)!)} = 15\,504 \tag{1}$$

Bearing in mind that in reality the product range sold in stores and the average *transaction width* are often much higher, the rapid increase of different product combinations is inevitable. To reduce the amount of itemset combinations, frequent itemset generation is applied.

Frequent itemsets can be defined as combinations of products that occur often in customer transaction data. Frequent itemset generation thus describes the process of efficient counting of the most commonly occurring items in transactions. In practice, frequent itemset generation involves finding all itemsets that satisfy the minimum support. *Support* can be expressed as *support count*, the number of times a certain item appears in the transaction data, or *support*, which refers to the number of transactions including the item divided by the total number of transactions. *Support* is also used as a measure of the strength of the association rule, and therefore it will be further explained in the later sections reviewing the

scoring methods of the association rules. (Chen et al. 1996; Zheng et al. 2001; Larose & Larose 2005; Tan et al. 2005).

The following subsections will explain the process of frequent itemset generation. First, the *Apriori principle* and the *Apriori algorithm* will be defined. A simplified depiction of the fundamental logic behind the *Apriori algorithm* will also be portrayed.

4.2.1 Apriori principle

The *Apriori principle* is an assumption adapted in the methodology of association analysis performance. It utilizes the *support* measure to reduce the number of itemsets explored during the frequent itemset generation. The principle can be described as follows: "if an itemset is frequent then all of its subsets must also be frequent" (Zaki 2000; Tan et al. 2005). Mutually, if an itemset is infrequent, then all of its subsets must also be infrequent. In practice this means that if itemset {A B} is infrequent, then all subsets i.e. {A B C} or {A B C D} derived must also be infrequent. This strategy can also be referred to as support-based pruning (Zaki 2000).

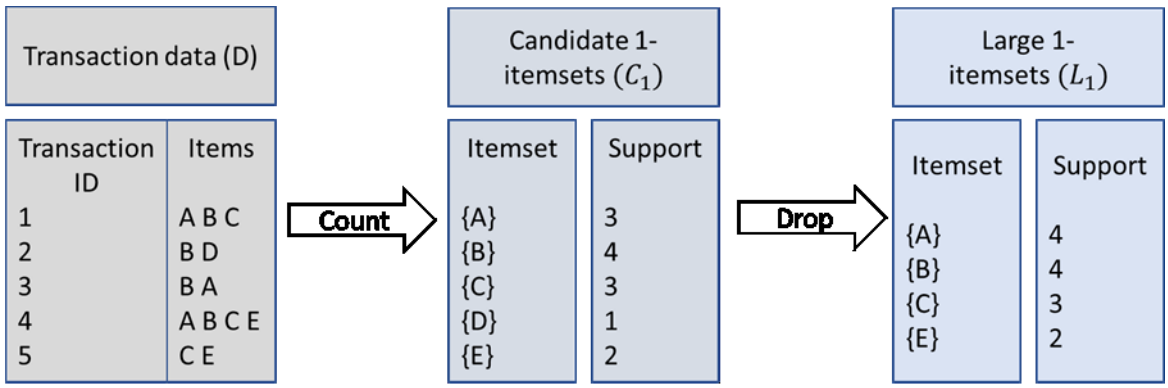
The *Apriori principle* acts as the basis for the *Apriori algorithm*, which can be utilized in frequent itemset generation (Chen et al. 1996; Zaki 2000; Zheng et al. 2001; Tan et al. 2005). The *Apriori algorithm* is the first algorithm in association analysis that pioneered the use of support-based elimination techniques to regulate the ascending increase of candidate itemsets (Zaki 2000; Tan et al. 2005). In all simplicity, the algorithm works iteratively by counting appearances of individual items in a transaction dataset, forming candidate itemsets based on a predetermined *support* level, and ends up with large itemsets up to any *k*-number of items still satisfying the support level constraint (Chen et al. 1996).

4.2.2 Apriori algorithm

The underlying logic behind the *Apriori algorithm* will be elaborated on in the following sections. This representation is extracted from Chen et al. (1996). In the first step (Figure 1), the algorithm scans all the unique transactions from the entire transaction database (D) and

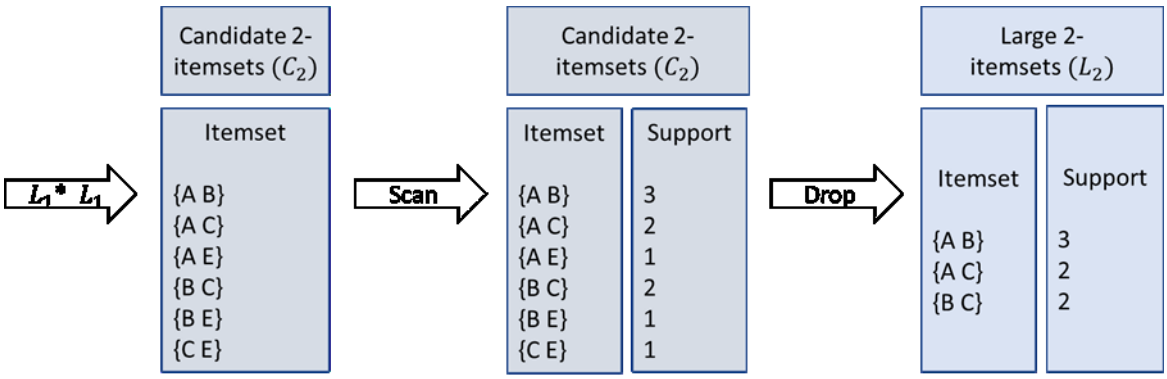
counts the occurrence of each individual item. Based on this count, the Candidate itemset C_1 is generated. As explained previously, the count here refers to the amount of times a certain item or itemset has occurred in the list of all transactions and is denoted in the "Support" column of C_1. Then, the algorithm reduces the number of itemsets by dropping those that do not satisfy the predetermined minimum *support* threshold, which is two in this case (or 40%). This results in the set of Large 1-itemsets L_1.

Figure 1: First stage itemset generation



In the second stage (Figure 2), the algorithm proceeds to discovering Large 2-itemsets (L_2) based on the principle that any subset of a large itemset must also have minimum *support*. The algorithm first generates a list of all the possible candidate 2-itemsets by an operation of concatenation of the Large 1-itemsets generated in the first stage (L_1). It then scans transaction dataset (D) and counts the occurrence of all Candidate 2-itemsets (C_2), denoted again in the "Support" column. Based on the same required *support* limit required, the Large 2-itemsets (L_2) are generated.

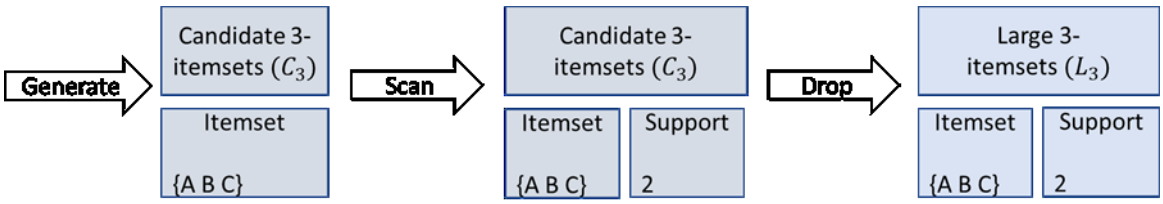
Figure 2: Second stage itemset generation



In the final stage, the Candidate 3-itemsets (C_3) are produced from the Large 2-itemsets (L_2) (Figure 3) by first identifying those 2-itemsets with the same first item. In this case, itemsets {A B} and {A C} would be chosen. Then the algorithm tests whether the 2-itemset consisting of their second items {B} and {C} in this case, forms a Large 2-itemset on its own. In this case it does, as it is included in the list in L_2. Since {B C} is a large itemset on its own, we know that all subsets of {A B C} are large itemsets, and therefore {A B C} forms a candidate 3-itemset in C_3. No other candidate itemsets from L_2 can be generated.

The algorithm again scans the transaction data (D) and counts the occurrence of itemset {A B C} and finds that it is represented two times in all transactions, and thus receives *support coun*t of 2. This satisfies the minimum threshold, and Large 3-itemset L_3 is generated. No itemsets needed to be reduced from the list of candidate itemsets, as itemset {A B C} was the only itemset present in the list. As there are no Candidate 4-itemsets to be deduced from L_3, the algorithm ends the process of discovering large itemsets.

Figure 3: Third stage itemset generation



4.2.3 Alternate approaches to frequent itemset generation

As the *Apriori algorithm* iterates through each individual transaction in the transaction dataset, it becomes evident that the computational resources required quickly increase. Given the reality that transaction data often consists of thousands if not millions of records of data, this restraint needs to be addressed when aiming to efficiently generate frequent itemsets. Also, the computational resources required in the frequent itemset generation are often significantly higher than those needed in the association rule generation, which is why alternate approaches have emerged.

Tan et al. (2005) propose two extensions to the *Apriori* approach to diminish the computational complexity related to the frequent itemset generation. Firstly, they suggest reducing the number of comparisons. The number of comparisons can be reduced by exercising one of the two strategies: storing the candidate itemsets or compressing the dataset. The candidate itemsets can be stored using the *Hash tree* in the support counting of the *Apriori algorithm.* The algorithm partitions the different itemsets into groups or buckets, and stores these in a *Hash Tree* model. This occurs while performing the *support count* associated with the individual itemsets included in each transaction, and results in concurrently "hashing" the itemsets into their appropriate groups. Instead of comparing each individual itemset in the transaction with every candidate itemset, the itemset in the transaction is only matched against those candidate itemsets that belong to the same bucket. This reduces the required amount of comparisons, and thus, the computational complexity of the algorithm. (Tan et al. 2005).

Another mode used to reduce the number of comparisons in the frequent itemset generation is to compress the data. Chen et al. (1996) propose the *FP growth algorithm* to be used to here. According to their definition "it *[FP growth algorithm]* generates all frequent itemsets satisfying a given minimum *support* by growing a frequent pattern tree structure that stores compressed information about the frequent patterns". Thus, repeated database scans can be avoided, and the generation of a large number of candidate itemsets can be heavily diminished.

Secondly, the overall number of transactions can be reduced in order to discourage the increase of computational resources required in frequent itemset generation. Vijayalakshmi and Pethalakshmi (2015) propose an alternative algorithm that reduces transactions in the transaction database by a repetition count approach (TR-RC). In this approach, the number of transactions can be reduced by decreasing the amount of similar transactions in the dataset. Vijayalakshmi and Pethalakshmi (2015) extend this method and suggest the use of another similar algorithm (CBTRA). This algorithm counts both the times a transaction is repeated in the database of all transactions and the relative size of the transaction. By applying this method, the number of similar transactions can be significantly reduced in addition to cutting down the unnecessary transactions (Vijayalakshmi & Pethalakshmi 2015).

The number of transactions can also be narrowed manually, by partitioning the transaction data into smaller proportions. It can be beneficial to fraction the data based on product groups or a certain time frame. For the division to be justified, the data losses must be accounted for so as to reduce manipulation of the final results of the analysis.

4.2.4 Other algorithms

Other algorithms commonly used in association analysis include *Direct Hashing and Pruning (*DHP*)*, *Charm* and *Closet*. The DHP algorithm is very similar to the *Apriori algorithm,* as it generates candidate *k*-itemsets also based on the previously constructed Large itemset L_{k-1}. DHP utilizes hash tables built in the previous scan to test the qualification of the *k*-itemset. However, instead of including all *k*-itemsets, the algorithm saves searching time in the hash tree by using filtering techniques. DHP both reduces the size of transactions and the amount of transactions in the database. (Chen et al. 1996; Zaki 2000).

The *Charm* and *Closet* algorithms on the other hand generate closed frequent itemsets. The closed frequent itemsets for association rules can be defined as subsets that represent the frequent itemsets without any loss of information. These algorithms enumerate the closed frequent sets, therefore eliminating the need for a bottom-up approach used by other common algorithms. This is an important feature when performing association analysis in

domains where the bottom-up approaches are not practically feasible due to long frequent itemsets. (Chen et al. 1996; Zaki 2000; Zaki & Hsiao 2007).

4.3 Association rule generation

Association rules govern the associations and causalities between disjoint sets of items A and B, and are expressed in the form of A → B. The most elementary form of an association rule involves two individual items A and B. These items can be products or product types, where A is called the antecedent and B is called the consequent, and thus, can be read as follows: if A then B. In the case of transaction data analysis, the association rule here could imply that when itemset or product A is purchased, itemset or product B is likely also purchased. (Srikant & Agrawal 1997; Zaki 2000; Zheng et al. 2001; Larose & Larose 2005; Tan et al. 2005).

In theory, the number of items *k* included in an association rule do not need to be confined to certain boundaries (Larose & Larose 2005; Tan et al. 2005). Thus, the association rules generated can potentially contain any number of items *k* ranging from 2 to *n.* In the case of transaction data analysis, *n* indicates the amount of all the different products that were sold during a certain time period where the data was collected.

As established in the previous section, frequent itemset generation is an efficient way to reduce the number of product combinations. However, frequent itemset generation might not be a feasible method on its own when the amount of items purchased by customers increase, which refers to the second major problem introduced in section 4.1. Let's now assume that there is no upper bound for the *transaction width* but instead, the customer transaction data can contain transactions with any number of items from the product portfolio including a total of 20 unique items.

From this transaction dataset the frequent itemsets are generated first. Suppose that the largest frequent itemset satisfying the predetermined minimum support threshold is a Large 10-itemset (L_{10}) and contains a total of 10 unique items. This one itemset alone can generate up to:

$$2^{10} - 2 = 1\,022. \quad (2)$$

different association rules, excluding rules that have empty antecedents or consequents. However, if the largest frequent itemset would contain 15 individual items, the itemset alone could generate up to:

$$2^{15} - 2 = 32\,768 \quad (3)$$

different association rules. This, combined with the knowledge that the example here is an extreme simplification, demonstrates the need to consider the number of items included in the frequent itemsets.

Association rule generation, or association rule discovery, identifies associations among the determined frequent itemsets given certain thresholds (Srikant & Agrawal 1997; Zheng et al. 2001; Larose & Larose 2005; Tan et al. 2005). In all simplicity, an association rule can be generated by subdividing itemset A into two non-empty fragments B and A – B, such that the rule B → A – B satisfies the predetermined boundaries. Consider for example a frequent itemset A {a,b,c}. From this itemset, a total of

$$2^3 - 2 = 6 \quad (4)$$

association rules can be extracted. These are :{a,b} → { c}, {a,c} → {b}, {b,c} → { a}, {a} → {b,c}, {b} → {a,c}, and{c} → {a,b} (Tan et al. 2005). Without setting fixed constraints for the rules to satisfy, all potential association rules will be generated. The rules as such do not imply associations or causalities as did not the individual transaction records suggest frequency of items. This is because the discovered patterns between itemsets may occur completely by chance.

Additionally, generating all possible association rules leads to a situation where most of the rules are in fact *irrelevant* or *uninteresting*. In order for an association rule to be relevant, it needs to be *strong*, i.e. have an adequate representation in the transaction dataset (Chen et al. 1996; Larose & Larose 2005; Tan et al. 2005). However, given a high score in the relative strength of the association rule does not imply that the rule is also *interesting.* In order for

an association rule to be classified as *interesting,* it not only needs to satisfy the objective interestingness measures similar to those evaluating *strength,* but also the subjective interestingness measures (Larose & Larose 2005; Tan et al. 2005).

This contributes to the third challenge related to association analysis, presented in section 4.1. By assigning various scoring methods to the generated rules enables detecting and evaluating the relative *strength* and *interestingness* of the rules (Chen et al. 1996; Tan et al. 2005). This is an elemental factor also in assessing the success of the algorithm in terms of its accuracy and thus, reliability. These scoring measures are exhibited in the following sections.

4.4 Association rule evaluation

The succeeding sections will first discuss the scoring methods related to the evaluation of the *strength* of the association rule. These methods involve objective measures *support* and *confidence.* The latter part of this chapter will examine the scoring methods associated with the *interestingness* of the association rule. These methods include both objective measures and subjective measures. The objective measures similarly include *support* and *confidence,* but also *lift, correlation* and *Piatetsky Shapiro.*

The subjective measures include the *unexpectedness* of the rule and the *actionability* of the rule. The subsections will be structured as follows: First, a definition of the scoring method will be provided. Next, the mathematical formulation of the scoring measure will be presented. Finally, the general practical applications of the rule will be briefly indicated.

4.4.1 Support

The *strength* of an association rule can be determined by calculating two different objective scores for the rule, the *support* and the *confidence*. *Support* can be subdivided in to *support count* and *support.* Similar to the definition given in chapter 4.2 “Frequent itemset generation”, where the occurrence of certain itemsets in the transaction data were calculated,

the *support count* of an association rule is defined as the number of times the association occurs in the transaction dataset. The support count can be formulated as follows:

$$Support\ count\ (A \rightarrow B) = a$$

where a describes the number of customers buying both product A and product B. With regards to the practical implications of this measure, it does not provide causality approximations, i.e. an estimation of the number of customers likely purchasing product B, having purchased product A. Nonetheless, the *support count* can be utilized as a minimum threshold for pruning the association rule candidates to reduce computational complexity, similarly to that utilized in the context of frequent itemset generation. The *support count* can also be used as a reference value to determine the feasibility of meeting the ROI requirements of a certain marketing campaign for instance. (Larose & Larose 2005; Tan et al. 2005).

The *support* of an association rule is often denoted as a percentage, and refers to the fraction of all transaction records for which the association rule applies (Tan et al. 2005). The *support* of an association rule can also be seen as an estimate for the probability of the association rule occurring, and is therefore a useful tool to prune those association rules generated by chance (Agrawal et al. 1993). *Support* is calculated as follows:

$$Support\ (A \rightarrow B) = \frac{Number\ of\ transactions\ with\ A\ and\ B}{Total\ number\ of\ transactions} = P\ (A \cap B)$$

Support can similarly be used as a pruning measure to reduce the computational power required to perform association analysis. Using *support count* or *support* to prune the association rules with low *support* often increases the statistical significance of the remaining rules. This is an important implication when dealing with scarce resources, for instance in determining the customer segments for which the marketing actions are targeted at. (Larose & Larose 2005; Tan et al. 2005).

However, using *support* in pruning the association rules produced might hinder the discovery of irregular, although important, behavioural habits of customers that may be concealed for example in negative dependencies.

4.4.2 Confidence

The other objective measure evaluating the *strength* of an association rule is c*onfidence.* The *confidence* of an association rule is also a percentage value and it determines the conditional probability of a consequent occurring given that an antecedent has occurred (Chen et al. 1996; Larose & Larose 2005; Tan et al. 2005). In the case of customer transaction data the *confidence* of rule A → B would describe the probability that a customer purchases product B given that they have added product A to their shopping cart. *Confidence* is calculated as follows:

$$Confidence\ (A \rightarrow B) = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A} = \frac{P\ (A \cap B)}{P\ (A)}$$

As described, *confidence* estimates the probability of an item being purchased given that another item has been selected. This has practical implications for instance in a sales situation, where the sales representative can promote the consequent (product B) for a customer who has chosen to purchase the antecedent (product A), given the high probability that the customer will respond to the promotion. With this in mind, association rules can be pruned to include only those with exceptionally high *confidence* scores.

Nonetheless, also *confidence* comes with its limitations when used on its own, as computing the *confidence* score for items with a very frequent consequent (product B) might provide misleading, or even paradoxical results. This is mainly due to the fact that the *confidence* score computes the joint probability $P\ (A \cap B)$ only against the probability of the antecedent $P(A)$. Assuming that the logic of statistical independence holds:

$$P\ (A \cap B) = \ P(A) * P(B)$$

in which case any deviation between $P\ (A \cap B)$ and $P(A) * P(B)$ would imply a statistical relationship. Given product B is very frequent, thus leading to a high deviation, the *confidence* score will therefore also be very high. This can occur with any product in the antecedent, provided a very frequent consequent, and with product combinations that are actually statistically independent, hence leading to contradicting results. (Chen et al. 1996; Larose & Larose 2005; Tan et al. 2005).

4.4.3 Lift

Support and *confidence* are used to measure primarily the *strength* of the identified associations, but they also contribute to the *interestingness* of the generated rules, since they can be used to prune infrequencies. Bearing in mind the practical limitations of both scores, the credibility of *support* and *confidence* in accurately measuring the *interestingness* of association rules is questionable, and for that reason, various other objective measures have been developed to better serve that purpose.

The first method to be discussed is the *lift* score. The *lift* score computes the joint probability $P\ (A \cap B)$ against the probabilities of both the antecedent $P(A)$ and the consequent $P(B)$. Thereby *lift* measures both the dependence of two items or itemsets and the nature of the dependence, i.e. whether the relationship between the items is attractive or repulsive. (Tan et al. 2005). *Lift* is calculated as follows:

$$Lift(A \rightarrow B) = \frac{Number\ of\ transactions\ with\ both\ A\ and\ B}{Total\ number\ of\ transactions\ with\ A * Total\ number\ of\ transactions\ with\ B}$$
$$= \frac{P\ (A \cap B)}{P\ (A) * P(B)}$$

Lift is considered an excellent measure of *interestingness,* since it provides an estimate of the statistical dependence between itemsets. The lift scores can be interpreted as follows:

Table 1: Lift score interpretation

Lift score	**Occurrence**	**Implication**
< 1	if A and B are negatively associated	Itemsets A and B are repulsive, they occur more seldom together than expected
= 1	if A and B are independent	Itemsets A and B are neutral, they occur randomly together in customer transactions
> 1	if A and B are positively associated	Itemsets A and B are attractive, they occur more often together than expected

The practical advantages of using *lift* as a scoring method include identifying and pruning *uninteresting* rules from the *strong* rules detected during the computation of *support* and *confidence*. As discussed, discovering *strong* rules among itemsets does not always imply that the rules are *interesting,* or that the dependencies are represented accurately. Therefore it might be useful to remove those association rules with a *lift* score close to one. What is more, *lift* can provide important insight into customer behaviour, as it recognizes also negative associations among itemsets. The itemsets with a *lift* score greater than one can potentially be used for cross-selling purposes, but with careful consideration, as will be explained in the next paragraph. (Larose & Larose 2005; Tan et al. 2005).

Nevertheless, also *lift* has its downfalls, namely in cross-selling contexts. This is because during the process of calculating the *lift* score, the conditional probability $P\ (A \rightarrow B)$, or *confidence,* between the two itemsets is lost. Accordingly, an itemset combination can have a high *confidence* score, implying high probability of the consequent resulting from the antecedent, but still result in a *lift* value less than one. Given the objectives of cross-selling, using *lift* on its own has limited applications as it does not provide information on the likelihood of itemset B being purchased given A has been chosen. (Brin et al. 1997).

4.4.4 Other objective measures

In addition to *support*, *confidence* and *lift*, also other objective measures commonly used to capture the interestingness of an association rule exist. These scores include the *Piatetsky-Shapiro (PS)* measure, *conviction,* and *correlation analysis.* The *PS* measure, also known as *leverage*, measures the difference between itemset A and itemset B occurring together in the transaction dataset and the expectation in a situation where itemset A and B are statistically dependent (Piatetsky-Shapiro 1993). The *PS* score is calculated as follows:

$$\begin{aligned} PS\ (A \rightarrow B) &= Support\ (A \rightarrow B) - \ Support\ (A) * Support\ (B) \\ &= P\ (A \cap B) - P(X) * P(Y) \end{aligned}$$

Similar to the interpretation logic of the *lift* score, a negative *PS* value indicates a negative relationship, a *PS* value of exactly zero indicates statistical independence, and a positive *PS* value suggests a positive relationship between the two itemsets. As with *lift*, the practical rationale of using the *PS* score helps in determining how many more combinations of itemsets A and B are actually sold in comparison to the expected sales of the individual itemsets. (Piatetsky-Shapiro 1993; Tan et al. 2005).

The *conviction* of an association rule was developed as a substitute scoring method to *lift,* to better capture the direction of associations. *Conviction* uses the consequent *support* and the computed *confidence* of the association rule to calculate the dependence of the consequent on the antecedent. In essence, *conviction* describes the frequency of the incorrectness of the association rule given that the items are independent of each other. (Piatetsky-Shapiro 1993; Tan et al. 2005; Hahsler 2019). The *conviction* is calculated as follows:

$$Conviction\ (A \rightarrow B) = \frac{1 - Support\ (B)}{1 - Confidence\ (A \rightarrow B)}$$

A high *conviction* value implies high dependence and a *conviction* value of 1 suggests statistical independence. More specifically, a *conviction* value equal to 1.2 signifies that rule A → B would be incorrect 20% more times if itemsets A and B were independent in comparison to the assumption that the items depend on each other (Hahsler 2019).

Another method to evaluate the interestingness of an association rule is to compute the *correlation* coefficient for the two itemsets. *Correlation analysis* is one of the most popular techniques used to study relationships between variables. The *correlation* coefficient for binary variables denoted as $\emptyset$ can be derived from the *support* scores of itemsets A and B as follows:

$$\emptyset\ (A \rightarrow B) = \frac{Support\ (A \rightarrow B) - Support\ (A) * Support(B)}{\sqrt{Support(A) * \big(1 - Support(A)\big) * Support(B) * (1 - Support(B))}}$$

Thus, $\emptyset\ (A \rightarrow B)$ can be interpreted as a normalized adaptation of the *PS* score demonstrated above, capturing the normalized difference between $Support\ (A \rightarrow B)$ and $Support(A) * Support(B)$. Consequently, $\emptyset = -1$ suggests a perfect negative relationship, $\emptyset = 0$ suggests independence and $\emptyset = 1$ suggests a perfect positive relationship. (Piatetsky-Shapiro 1993; Tan et al. 2005).

4.4.5 Unexpectedness and actionability

Although the objective scoring measures offer useful and straightforward methods to gauge the *strength* and *interestingness* of association rules, they are not considered sufficient when aiming to identify previously unknown patterns in customer buying behaviour (Tan et al. 2005). Take for instance an association rule (A → B) classified as *strong* and *interesting* based on its high objective scores: *support*, *confidence* and *lift,* where A refers to item "Hot dog" and B refers to item "Sausage". Despite being objectively *interesting*, the rule can be determined subjectively *uninteresting,* since it reveals quite an obvious linkage; a customer buying hot dogs often also buys sausages.

Now consider an association rule (A → B) that is *strong*, *interesting* and *unexpected*, thus satisfying objective scoring methods and not revealing any obvious pattern. In this example, let's base this assumption on the fact that the rule has a high *lift* score. This could mean for instance that customers often purchase "Hot dogs" together with "Beer". However, the association rule and its dependency estimation on its own is of no use, if it does not fit the company's objectives, that is to say, if the rule can't be acted upon. For instance, a rule with a high *lift* i.e. "Hot Dogs" and "Beer" is not *actionable* if the company plans to quit selling "Beer". Hence, *actionability* of the rule depends on the objectives of the company, i.e. what they want to achieve with the association analysis. Therefore, in order for the association be also subjectively *interesting*, it needs to be *actionable*. (Tan et al. 2005; Geng & Hamilton 2006).

In order to select the most appropriate pruning and scoring methods, it is extremely important to understand the fundamental logic behind each method, as well as their practical propriety. No measure can be undeniably proclaimed as superior, because of their intention-specific nature. An advisable course of action would be to weigh all measures concurrently. Table 2 presents a summary of the scoring methods.

Table 2: Summary of scoring methods

Scoring method	Description	Practical use	Limitation
Support count, Support	How frequently the association occurs in the transaction dataset?	Prune infrequent rules to reduce computational complexity Determine coverage of mass marketing actions	Pruning rules based on support might hinder finding interesting patterns that are infrequent
Confidence	What is the probability that product B is purchased given that product a has been purchased?	Promote product B when customer has bought product A	Using only confidence might provide misleading or paradoxical results given a frequent consequent
Lift	Are itemsets statistically dependent, and if so, are they positively or negatively associated?	Prune uninteresting rules from those with high confidence and support Increase sales by proposing itemsets that are often bought together Find interesting customer insight	Limited cross-selling use because it does not represent causality between buying itemset A and B
Conviction	How often does the rule make an incorrect prediction if the association between the items occurs by random chance?	Prune uninteresting rules Increase sales by promoting itemsets that are often bought together Find interesting customer insight	
Piatetsky-Shapiro, Correlation coefficient	What is the absolute/normalized difference between itemset A and B occurring together and them occurring individually?	Prune uninteresting rules Find interesting customer insight	Limited cross selling use because it does not represent causality between buying itemset A and B
Unexpectedness	Was the association expected?	Find interesting customer insight	Thorough product knowledge required
Actionability	Can the rule be acted upon based on the company's targets?	Apply association rules in business environment	Analysis method comprehension required

4.4.6 Association rule generation in Apriori

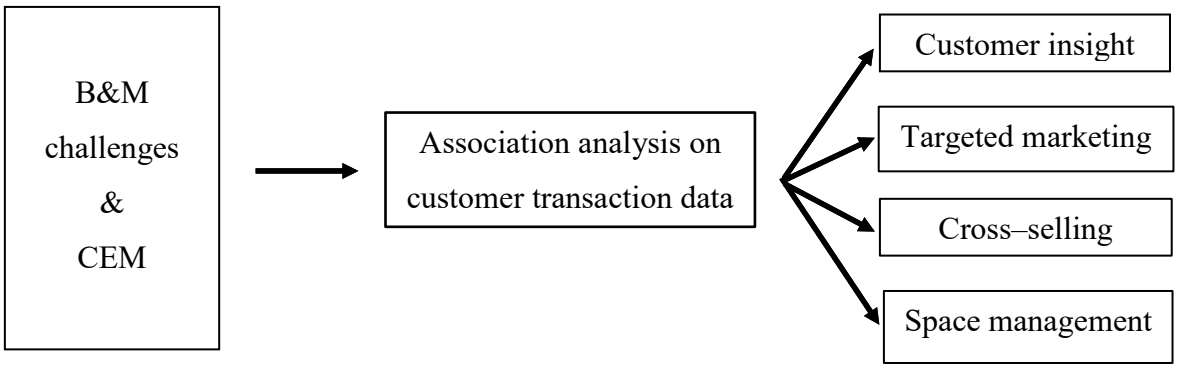
Like in frequent itemset generation, the *Apriori algorithm* uses a level-wise method to generate the association rules. The levels in frequent itemset generation represented the number of items belonging to that candidate or large itemset i.e. Large 3-itemset (L_3) represented level three and included three items. In association rule generation each level corresponds to the total amount of items belonging to the rule consequent. The algorithm then moves from the first level, i.e. extracts all high *confidence* rules with one item in the consequent, and then uses these rules to generate new candidate rules. (Larose & Larose 2005; Tan et al. 2005).

The only difference between the logic of the algorithm when employed in association rule generation compared to frequent itemset generation is that no additional scans of the transaction data is required to compute the *confidence* of the candidate rules. Instead, the *confidence* is computed from the *support counts* generated from the frequent itemset formation. (Larose & Larose 2005; Tan et al. 2005).

5 THESIS FRAMEWORK

The thesis framework is presented below in Figure 4. The framework summarizes the background circumstances that establish the motives behind this thesis. In the centre is presented the core element in this research, which is utilizing association rule mining to analyze customer transaction data. The outcomes of this research are portrayed on the right.

Figure 4: Thesis framework



Two major occurrences have motivated carrying out this research. Firstly, it is the current challenges burdening the B&M store profitability prospects that have sparked the need to re-evaluate the in-store customer experience. These issues are manifested in various retail markets, and have shown to influence also paint and paint supply stores globally. The initial investment required to set up a physical store may not add up to much. In contrast, succeeding in acquiring a solid customer base and managing to outperform budgeted net sales on a yearly basis is a complex mission.

Secondly, it is the case company's strategic re-organization and progression towards successful customer experience management that have triggered interest towards this topic of study. This initiated concern over examining the theoretical construct itself as well as the interdependencies between the theory and the contribution of this research. The literature review defined customer experience and how it has developed into customer relationship management and customer experience management. It was also examined to what extent customer transaction data plays a role here.

Further, it was established that customer transaction data can be used to gain interesting insight on customer purchase behaviour. This information may reveal details into customer preferences that would not be uncovered by alternative means, say, qualitative customer interviews. Next, association analysis techniques were consolidated with sales increasing strategies. The sales increasing strategies include targeted marketing, cross-selling and space management.

6 DATA AND ANALYSIS PROCESS

This chapter forms the beginning of the empirical part of this thesis. This part describes the data used, as well as the processes used in cleaning and preparation of the data. Following this, is the association analysis process.

6.1 Data description

This section describes the data used in this thesis. The data was collected from one of the case company's oldest paint stores, and hence, provided the largest depository of customer transaction data. The original data was in PDF format, and consisted of 11 730 customer transaction receipts that were extracted from the POS system. The transactions included all transaction items that actualized during year 2018, from 01.01.2018 to 31.12.2018.

First, to ensure protecting the case company's business and the privacy of their customers, it is essential to undertake action to manage the presentation of the original data and the results derived from there. Any potential customer identification data is concealed in the transaction receipts. Also company employee reference information as well as product details are removed from the original data formats. In addition, product names relevant to the presentation and discussion of the analysis results are assigned new names. These names represent vague equivalents of the original product names, therefore ensuring confidentiality while providing a measure of interpretability. The following sections will describe the data and its attributes and take a look at each step taken to prepare the data for the association analysis.

One transaction receipt (see example of an authentic receipt attached in Appendix 1) consists of the following information:

Table 3: Transaction receipt data

Field	**Description**
Fakturadatum	Date of transaction
Kundnummer	Customer ID
Fakturanummer	Transaction ID
Kund	Customer information
Er referens	Customer reference
Vår referens	Company reference
Varunr	Product ID
Varunamn	Product name
Anmärkning	Product count
Följesedel total	Total amount in Swedish Crowns

As depicted in Appendix 1 and the table above, the transaction receipt includes customer information such as the customer ID and customer name. However, it is important to note here that this customer-specific information was missing in the majority of the transaction receipts or was partially lost during the data format transfer in cases where the "Customer reference" was stored in the same column as the product details.

This information loss is not grave when considering the scope and objectives of this thesis, although it is essential to be aware of the repercussions that follow. Since the customer identification details are missing, the exclusivity of the transaction, and therefore the uniqueness of the customer cannot be confirmed with certainty. This indicates that a certain customer can have potentially performed the identical transaction during multiple occurrences, which why the transactions do not automatically belong to unique customers. Nonetheless, this pattern does not occur frequently in this specific dataset, which was revealed after a brief scan for duplicate transactions. Therefore, it will be assumed that the individuality assumption holds.

6.2 Data cleaning

First, all customer transactions were merged into one single PDF file using Docs.Zone. Docs.Zone is a web-based file conversion software that enables merging independent files within a uniform file format and translating files from one file format to another ("Docs.zone convert files" 2019). This master PDF was then converted into Excel format (Appendix 2). This however resulted in a significant number of ambiguous rows and columns including null or indeterminable values, and thus, required in depth querying.

The sheet was further edited in Power Query editor. Power Query is an extension available for Microsoft Excel and Power BI desktop. The tool allows to edit large datasets by searching specific sources, changing the data types and merging rows and columns. First, all null cell values, rows and columns were deleted. Then, all single-item transactions were removed, as they are not applicable in association analysis. Next, the column containing the product name was concatenated into a single list, each row corresponding to products bought within one customer transaction. This list was merged with the corresponding Date values and finally exported back to Excel.

6.3 Data preparation

After the initial data cleaning, some auxiliary steps were taken in order to proceed with the association analysis. All individual items bought were extracted from the list of customer transactions by merging the transactions into a single cell, detaching separate products by comma-delimitation, and converting the text into individual rows with Excel's Text to Columns Wizard. The duplicate item names were removed and the list was transposed into column headings. These headings were then combined with the original list of customer transactions into a 7291*1224 size array, the former representing the customer transaction rows and the latter the column headings comprised of the product names.

Lastly, TRUE and FALSE values were computed for each customer transaction. Here value TRUE declares existence of a certain product in a customer transaction and FALSE indicates

the absence (Appendix 3). This discretionary step was taken to modify the dataset into a format that is easily transferrable and analyzable in the programming tool.

The following sections will describe how the association analysis was performed on the customer transaction data. Firstly, the technical analysis appliances will be reviewed. Then, it will be described how the process of association analysis is applied to this specific research, considering the resources available. The last two sections describe the frequent itemset generation and association rule generation.

6.4 Analysis tools

As discussed in the methodology part, different models and approaches to association analysis exist, the most common being the adoption of machine learning techniques and algorithms. Numerous user-friendly applications and software have been developed for these purposes, but the method used in this thesis combines programming, data analysis and machine learning packages to produce the desired output. This section will describe the technical tools used including the programming language, user environment and the different data analysis packages.

6.4.1 Programming language and user interface

This thesis uses Python programming language within Anaconda's Jupyter Notebook user interface. Python was chosen for its interpretability, comprehensiveness and large standard library. Two environments for Python 2.7 were set up to enable the transaction data analysis: the Jupyterhub environment, and the Jupyter Notebook user environment. In addition, Anaconda 2019.10, an open-source platform including data science packages, models and visualization tools, was installed.

Jupyter Notebook is an open-source web application allowing its user to create and distribute live code, numerical equations, data visualizations as well as narrative text. It is a flexible tool used in data mining and visualization, statistical model development and testing, machine learning, optimization, simulation and so forth ("The Jupyter Notebook" 2019).

6.4.2 Packages

Previously, Python programming language has been used specifically for data preparation and querying, and not so much for data analysis purposes. However, the development of various toolkits and libraries has enabled the implementation of the entire data analysis workflow within the Python environment, without having to switch to a domain specific programming language. The primary toolkits or packages used in this thesis are Pandas, Numpy, Sklearn, Matplotlib and Mlextend.

Pandas provides high-performance tools for data structuring. These tools include for instance the DataFrame object for data remodeling, functions for interpreting and extracting data between different file formats, flexible grouping, merging and joining of data sets, and intelligent handling of missing data values. Numpy is a core package for scientific computing, including a powerful array object and useful tools for linear algebra. Numpy is the underlying foundation for which machine learning packages such as the Scikit-Learn and Matplotlib are built on. Scikit-Learn, or Sklearn, provides different machine learning algorithms for classification, regression, clustering, and model selection problems, whereas Matplotlib is used for visualization and plotting. Finally, Mlextend is an extension used in this thesis for the association analysis. (DataFlair Team 2019; “Github: mlxtend” 2019).

6.5 Application of association analysis

This section focuses on describing the case specific application of the methodology described in chapter 4. Association analysis is adopted due to two major reasons. Firstly, from the case company’s perspective, it is desirable to utilize available customer data in a manner that is novel to the company. Considering that the available data is transaction data without historical purchase information or customer identification references, finding associations between product groups is the appropriate solution. Secondly, association analysis has the potential of uncovering inexplicit behavioural patterns of customers, that may not be revealed by other means, for instance by conducting customer interviews. Besides, association analysis is often connected to targeted marketing, cross-selling and

space allocation decisions of companies, as disclosed in the literature review, which is why it provides important practical managerial implications as well.

This thesis adopts the *Apriori algorithm* as the fundamental method used in association analysis. This approach is feasible due to the modest processing times of the algorithm coupled with the confined set of customer transactions.

Considering the diversity in the transaction data resulting from the wide product range sold, the number of items need not be restrained in frequent itemset generation. This is because the frequent itemsets generated based on the chosen *support* level do not produce an overwhelming number of itemsets to be accounted for. Additionally, to not obstruct identification of interesting, yet infrequent customer behaviour patterns, no strict pruning thresholds will be set for association rule generation.

With regards to the subjective scoring of the association rules, the *unexpectedness* and *actionability*, the following assumptions hold. Both company representatives and product information data will be referred to when addressing *unexpectedness* of the associations. *Actionability* is evaluated based on whether the rule revealed interesting customer insight or can be utilized to increase sales.

Recall that sales in this context can be increased by employing targeted marketing, cross-selling or space management tactics. The former two tactics are similar in nature, but differ in the sense that cross-selling tactics here refer to concrete sales arguments or recommendations performed at existing customers, whereas targeted marketing can be applied to any existing or potential customer. Space management strategies in this case focus on product placement decisions that aim to increase the probability of impulse purchases rather than improving operational efficiency.

6.6 Frequent itemset generation

First, the customer transaction data array was imported to Jupyter Notebook from Excel as Pandas DataFrame object. The column headings denote the product names, ranging from

"Product 1" to "Product 1224" (Appendix 4). The original DataFrame object includes also the dates of the transactions in no particular order.

After importing the data, the date column was removed and the new DataFrame was stored as "records". As preliminary exploratory data analysis, the *Apriori algorithm* in the Mlextend package is used to generate the most frequently bought individual items. Note that a minimum support of 0.02, or 2% is used here, which corresponds to 146 transactions. This implies that in order for an item do be determined frequent, it needs to exist in at least 146 customer transactions. Varying minimum support thresholds were examined between range [0.001 : 0.2], before arriving at 0.02, which had a reasonable processing time relative to the output produced. Furthermore, the itemset length was adjusted to < 2 in order to include only individual items, as opposed to those itemsets including several products, generated for association analysis purposes, described next.

The code was then modified by reducing the required minimum *support* to 0.015, or 1,5% (Appendix 5). This was a necessary step in order to provide a sufficient amount of itemsets to be analyzed, since setting the itemset length to >= 2 reduced the number of transactions satisfying this constraint. To clarify this point, the minimum *support* threshold set at 1,5% combined with the item length constraint of >=2 implies that the output generates those product combinations that exist in 1,5% of all transactions, or 109 transactions in total, with a minimum of two items purchased.

As described in section 4.2, the *transaction width,* and therefore also the potential length of the frequent itemset, greatly influence the total number of association rules generated. Consequently, it is often desired to inhibit the itemset length already during the frequent itemset generation phase. However, the transaction data used in this thesis comprises of customer transactions that differ a lot from each other due to the wide product portfolio. Furthermore, the *transaction width* is relatively narrow even with the largest transactions. Combining these features with the minimum *support* threshold used, it was not necessary to set an upper bound to the item length of the frequent itemsets generated.

6.7 Association rule generation

Using the frequent itemsets generated in the previous step, the association rules were produced using the association_rules() function from the Apriori library (Appendix 5). Also here, various pruning levels were tested in order to observe the effect that the thresholds have on the number of association rules generated. The rules were first pruned based on the *confidence* of the rule. Recall that the *confidence* of rule A → B describes the likelihood that a customer purchases product B given they have chosen to purchase product A.

Altering the minimum *confidence* threshold between the range [0.50 : 0.99] only slightly reduced the number of association rules generated from 122 to 37. Given that rule A → B can have a high *confidence* if itemset B is very frequent, the *lift* scores for the association rules were also generated. Different threshold were also altered here, but the output of generated rules did not exceed the 122 generated earlier.

Recognizing that the total amount of association rules only added up to 122, and the fact that a majority of the rules were redundant, it was not a prerequisite to prune the rules further in order to continue with the association analysis. Instead, all the rules will be presented and analyzed in the further sections. This strategy also supports the approach discussed in section "Application of association analysis" on the chosen pruning strategies and their implications on identifying irregular customer buying habits. It was determined that a low pruning threshold would be chosen so as to not hinder identifying interesting customer insight.

7 RESULTS AND EVALUATION

This section will exhibit and evaluate the results of the association analysis as per the circumstances and course of action described in the previous sections. The results of the frequent itemset generation will be disclosed first, followed by the generated association rules. The findings will be evaluated in light of the scoring methods and the objectives of this thesis.

7.1 Frequent itemsets

Firstly, the most frequently purchased individual items were queried from the array of transactions using a minimum *support* of 2%. Note that the transaction DataFrame object is no longer the original customer transaction data consisting of all purchases during the one year time span, and therefore does not contain those transactions with only one single item, as these transactions were deducted during the data cleaning process described in section 6.2.

The most common products in descending order of *support* are presented in Table 4. According to this analysis, “Putty A” is the most frequently bought product, and it exists in 656 customer transactions, thus having a *support* of 9.01%.

Table 4: Frequent items

Item	**Support Count**	**Support**
Putty A	656	9.01%
Sealer	647	8.88%
Filler A	589	8.08%
Wall paint A	401	5.50%
Tape A	356	4.88%
Waste bag	336	4.61%
Ceiling paint	322	4.42%
Wall paint B	292	4.00%
Paint brush A	251	3.44%
Tinting product	234	3.21%

Next, the most frequent item combinations were extracted. Table 5 below summarizes the results. According to this analysis, the most common itemset is "Putty A" and "Filler A", implying that these products are most frequently bought together. They occur in a total of 564 customer transactions, and thus have a combined support of 7.74%.

Table 5: Frequent itemsets

Itemset	**Support Count**	**Support**
Putty A, Filler A	564	7.74%
Sealer, Wall paint B	252	3.46%
Putty B, Putty C	164	2.25%
Wall paint A, Lacquer	141	1.93%
Roller, Plastic can	126	1.73%
Roller, Glove	125	1.71%
Glove, Plastic can	125	1.71%
Roller, Glove, Plastic can	125	1.71%
Tape A , Tape B	122	1.67%
Filler B, Filler C	115	1.58%
Primer, Wall paint C	111	1.52%

7.2 Association rules

The association rules were then generated from the frequent itemsets. As disclosed in the preceding chapter, the total number of association rules added up to 122, when setting moderate minimum thresholds for *confidence* and *lift*. The full list of association rules can be found in Appendix 6. This includes the item name of the antecedent and consequent, the *support* of the antecedent and consequent, the *support*, *confidence* and *lift* scores of the association rules, as well as the *PS* value and *conviction.* For a detailed description of each scoring method refer to the end of section 4.4.5.

From these rules, a sanity check similar to that of the frequent itemset presentation was performed (Appendix 7). However here the product specific details (ie. “Base 1”) were not excluded, but instead, removed from those rules where the product detail was already included in the rule antecedent. This step was performed manually after applying the *Apriori algorithm* to generate the association rules, to ensure all items and product characteristics (ie. “Base 1”) were included to mitigate the risk of losing valuable information.

This additional step also contributes to the subjective scoring of association rules. By removing those association rules with recurring items or product information in the antecedent and consequent, preliminary unexpectedness of the produced rules can be assured before going further. Consider rule A → B, where item A stands for product “House paint base 1” and item B stands for “Base 1”. This rule would imply that item A in with base paint number 1 suggests a probability of base paint 1 occurring also in the consequent, which is rather obvious given the item name itself already includes the consequent.

7.3 Evaluation

As has been affirmed, the association rules generated strictly depend on the predefined metrics, or evaluation scores. Furthermore, the accuracy, and therefore reliability of the generated association rules vary, contingent on the chosen evaluation method. Hence, it is desirable to examine the association rules through various perspectives.

The following sections will look into the association rules in more detail. The rules will be presented and assessed based on the evaluation scores, their practical propositions, as well as their limitations discussed earlier. In cases of redundant rules, i.e. A → B and B → A, that have scored equal results in the ranking method in question, only one rule of the pair will be presented. Refer to Table 13 for a summary.

7.3.1 Support

Recall that the *support* of an association rules refers to the frequency that the rule occurs in the transaction data. The rules with the highest *support* are as follows:

Table 6: Highest support rules

	Rule	**Support**
1.	Putty A → Filler A	7.74%
2.	Sealer → Wall paint B	3.46%
3.	Wall paint A → Base 1	2.84%

These rules represent the fraction of the rules that exist most frequently in the customer transactions. For instance, rule Putty A → Filler A applies to 7.74%, or 564 of all customer transactions.

7.3.2 Confidence

As outlined earlier, *confidence* of an association rules aims to answer the following question: what is the probability that product B is purchased given that product A has been purchased? The rules with the highest *confidence* are as follows:

Table 7: Highest conidence rules

	Rule	**Confidence**
4.	Glove, Roller → Plastic can	100%
5.	Glove, Plastic can → Roller	100%
6.	Plastic can, Roller → Glove	100%

In contrary to the most frequent rules extracted based on their high *support,* these rules represent causalities between itemsets, hence, the order of the itemsets is crucial when interpreting the rules. Therefore, the rules above cannot be considered redundant, since they portray different information. Rule number four states that a customer purchasing product "Glove" and "Roller" purchases product "Plastic can" with a 100% likelihood. Rule number five on the other hand declares a slightly different causality between the itemsets as the consequent changes.

However, as explained earlier, a high *confidence* score of an association rule can sometimes provide misleading, or even paradoxical results. This occurs in the case where the consequent item or itemset has a high *support* in the transaction data. Besides, a statistical probability of 100% in any given circumstance is often erroneous, which is why a re-evaluation of the above presented rules is justified.

Let's first consider the potential paradoxicality of the rules. According to the frequent itemsets presented in the previous sections, the most common individual items in the transaction data were:

Table 8: Frequent itemsets (2)

Item	**Support Count**	**Support**
Putty A	656	9.01%
Sealer	647	8.88%
Filler A	589	8.08%
Wall paint A	401	5.50%
Tape A	356	4.88%
Waste bag	336	4.61%
Ceiling paint	322	4.42%
Wall paint B	292	4.00%
Paint brush A	251	3.44%
Tinting product	234	3.21%

The rules generated based on their *confidence* scores described above do not include any of the frequent itemsets. This implies that the results are trustworthy in that respect. However, the rules only include three distinct products, and have *confidence* scores of 100%, which is why a new set of rules is presented below.

Table 9: Highest confidence rules(2)

	Rule	**Confidence**
7.	Filler B → Filler C	94.26%
8.	Filler C → Filler B	93.50%
9.	Putty B → Putty C	86.32%

Observe that the *confidence* scores are already more reasonable, i.e. do not indicate 100% probability of the consequent occurring given that the antecedent has occurred. In addition, the number of items included in the rules has generally also reduced. Regardless, analysing this set of association rules already takes one step forward into identifying interesting causalities between items. For instance, the first rule implies that a customer buying “Filler B” very likely also buys “Filler C”. Also, a customer buying “Filler C” likely buys “Filler

B". In the case of these two rules, the order of items does not significantly change the interpretation of the rule as the *confidence* score does not change much.

7.3.3 Lift

The following section will evaluate the association rules based on their *lift* scores. The *lift* value indicates statistical dependence between two itemsets, whether it be positive or negative, but does not suggest causality. The top rules based on their statistical dependence are as follows:

Table 10: Highest lift rules

	Rule	**Lift**
10.	Filler C → Filler B	55.88
11.	Putty B → Putty C	38.37
12.	Wall paint C → Primer	32.76

All above presented itemsets seem to be very dependent on each other. A positive *lift* suggest positive dependence. In fact, all the association rules that were generated have a positive *lift,* denoting that none of the associations between the itemsets in this customer transaction data are repulsive, i.e. buying a certain item would decrease the probability of buying another item. However, it is possible that these types of associations were discarded already in the frequent itemset generation process, since no strict pruning thresholds were set for the association rule generation.

Recall that *lift* measures the performance of an association rule by computing the ratio of the predicted response divided by the average response. Rule 10 here suggests a ratio of 55.88 to 1 of the association rule predicting the consequent of the rule given the antecedents, compared to a random guess.

As professed earlier, extracting association rules by their *lift* values can provide interesting customer insight. Yet, this interesting insight might be rather infrequent, and thus, is lost already during frequent itemset generation, even with relaxed pruning thresholds. For

instance, the rules generated based on the frequent itemsets subject to *support* constraint 1.5%, do not include any *lift* values below 1. This would imply that none of the items purchased are repulsive in nature. Whether this is the actual case, or whether these negative associations were pruned in the early stages of the association analysis will be evaluated in the following sections.

7.3.4 Sensitivity analysis of support threshold

Let's first reduce the support threshold from 1.5% to 0.5%. This generates a total of 256 frequent itemsets, compared to the 38 generated when using *support* threshold 1.5%. From these itemsets, the new association rules were produced. This in turn delivered a total of 848 associations. The top and bottom rules based on *lift* are listed below, excluding redundant rules:

Table 11: Highest lift rules (2)

	Rule	**Lift**
13.	Paint brush B → Paint brush A	169.56
14.	Filler A, Filler B → Filler C, Putty A	151.90
15.	Roller product → Roller	70.11

Table 12: Lowest lift rules

	Rule	**Lift**
16.	Sealer → Putty A	0.76
17.	Filler A → Sealer	0.79
18.	Filler A, Putty A → Sealer	0.79

We observe a significant increase in the calculated *lift* scores for this new set of rules. Also here it is important to combine the *lift* values with the *confidence* of the consequent(s) occurring, given the antecedent(s), in order to make justified recommendations.

More notably, we now also detect negative associations between product combinations indicating repulsion. The first rule in the list of rules with lowest *lift* scores suggests that the rule is actually 0.76 times more often incorrect, compared to a random guess, which in this case implies selecting two random items from the transaction data and combining them in one individual transaction. In other words, it is 24% more unlikely that items “Sealer” and “Putty A” are purchased together, compared to any randomly chosen set of items with no prior statistical relation.

Interesting customer insight can also be found when looking at alternative objective measures, namely the *conviction* and *leverag*e of the association rules. The *conviction* is similar to *lift*, calculating how often a rule makes an incorrect prediction when the itemsets present in the rule actually only occur by chance. In contrary to *lift*, *conviction* also utilizes the *confidence* of the association rules, therefore assessing the direction of the rule, and hence, causality between itemsets. As is exhibited in the full list of generated association rules attached in Appendix 6, the *lift, conviction* and *leverage* scores of the association rules are correlated.

7.3.5 Unexpectedness and actionability

A summary of all rules evaluated in this chapter is presented in Table 13, with a subjective evaluation of the *unexpectedness* and *actionability* of the rule. To assess the *unexpectedness*, each item name was queried from the company database in order to access the item-specific details, including the properties and functionality of the product. The rules were declared as expected, if the rule antecedent and consequent are related to each other with regards to their functionality. In addition, a case company representative responsible for the paint store network in Sweden, was consulted. A list of product names that were identified during the process of association analysis was delivered, with a request to fill in the product descriptions according to his best knowledge. It was also asked whether the product identification details describe an individual product name, or a categorical label, that acts as a collective term used to describe multiple different products.

In addition, rule number three and nine were also determined expected, because the rule consequent actually describes the categorical label of the rule antecedent. These rules are

not *actionable* either. The rest of the rules were determined *unexpected,* if these properties did not hold. The *actionability* of the rule is also stated based on the objectives to either find interesting customer insight or propose sales increasing strategies.

Table 13: Summary of association rules

Section	#	Rule	Unexpected?	Actionable?
7.2.1	1	Putty A → Filler A	No	No
7.2.1	2	Sealer → Wall paint B	Yes	Yes
7.2.1	3	Wall paint A → Base 1	No	No
7.2.2	4	Glove, Roller → Plastic can	Yes	Yes
7.2.2	5	Glove, Plastic can → Roller	Yes	Yes
7.2.2	6	Plastic can, Roller → Glove	Yes	Yes
7.2.2	7	Filler B → Filler C	Yes	Yes
7.2.2	8	Filler C → Filler B	Yes	Yes
7.2.2	9	Putty B → Putty C	No	Yes
7.2.3	10	Filler C → Filler B	Yes	Yes
7.2.3	11	Putty B → Putty C	Yes	Yes
7.2.3	12	Wall paint C → Primer	Yes	Yes
7.2.4	13	Paint brush B → Paint brush A	Yes	Yes
7.2.4	14	Filler A, Filler B → Filler C, Putty A	Yes	Yes
7.2.4	15	Roller product → Roller	No	Yes
7.2.4	16	Sealer → Putty A	No	Yes
7.2.4	17	Filler A → Sealer	Yes	Yes
7.2.4	18	Filler A, Putty A → Sealer	Yes	Yes

These findings confirm the indisputable value of quantitative data analysis in terms of revealing concealed behavioural patterns. This revolutionary insight on customer purchase preferences has not been detected when conducting qualitative customer interviews. Furthermore, the only prior study on sold products has included items manufactured and owned by the case company, disregarding all other products and brand sold in the store, providing biased results.

The association rules derived are eye opening in many respects. First, the large support of ancillary products, i.e. Glove, Roller, Plastic can, Paint brush A and so forth, is rather astounding compared to the representation of the products that are manufactured by the case company and thus, considered a core focus. What's more, the results contradict the pre-assumed product categorisation where certain product brands are considered substitutes to competing brands, and therefore, the hypothetical likelihood of those products being bought together is rather minimal. See for example rule 14. Here products "Filler A" and "Filler C" are actually being bought together, although considered very similar in terms of functionality.

Meanwhile, the results show rather interesting evidence concerning the negative associations. Not only were these repulsions previously unidentified, but they do not seem to follow any certain logic, i.e. substitution. Perhaps one of the most interesting finding is rule 14 and 18, one depicting extremely strong attraction and the other showing repulsion between the products. When looking at the products included in the rule, they seem to be almost identical. Both include Filler A and Putty A, but when these products are combined, the likelihood of Sealer being purchased is significantly reduced. These and other findings will be discussed in depth in the next chapter.

8 DISCUSSION

This chapter will firstly revisit the identified research questions and aim to answer them based on the results presented in the previous chapter. Conclusive remarks follow, providing an evaluation of the success of this research relative to the capability to answer the identified research questions and the achievement of the pre-determined objectives.

8.1 Research questions

The first research question was as follows:

Which product combinations are frequent?

To answer this question, the frequent itemsets were generated from the customer transaction data and the top three results were Putty A & Filler A, Sealer & Wall paint B and Putty B & Putty C.

The itemset *support* levels in general were relatively low, corroborating the assumption of fluctuating customer purchase behaviour and the wide product range. Considering that the one item transactions were excluded from the data prior to analysis, reducing the sample size of the data from roughly eleven thousand to seven thousand, this observation is reinforced even to a greater extent.

The average *transaction width* was observed to be rather narrow, which also contributed to the frequent itemset generation. When analyzing transaction data with narrow *transaction width* the probability of detecting meaningful product combinations diminishes, as the overall number of different possible product combinations is reduced. This is important to keep in mind when evaluating the feasibility of the transaction data to be analyzed. However, a narrow *transaction width* is particularly useful when coping with scarce computational resources.

The resulting itemsets differed in the sense of their *unexpectedness*. The occurrence of certain product combinations together in customer transactions was rather self-evident, particularly in cases where causal relationships exist within the functionality and use of the itemset. For instance, some items that were frequently purchased together require application of the other prior to using the other i.e. primer paint and topcoat.

What's more interesting, is the categorical allocation of the most frequent itemsets. Majority of the products are actually ancillary products considered as daily necessities of the target customer group: professional painters. These products are often not considered a strategic priority, or even a topic of interest for the case company. Only three of the most common product combinations actually include the case company's core products, which are paints. In fact, these three wall paints are the only products included in the list of frequent itemsets that the case company manufactures, whereas the other products are purchased from suppliers. This highlights the importance of supply chain management and particularly the significance of various supply contracts.

Another noteworthy observation from the frequent itemsets is that some of the product combinations, i.e. that including items "Putty B" and "Putty C", comprise of products with very similar properties, and therefore assumed to be substitutes to one another. Indeed, many of these itemsets include two competing products. This detected purchase behaviour pattern, indicating that these products are commonly bought together, actually suggests complementarity rather than substitution. This is particularly interesting also because it contradicts the general perception of the customer base being loyal to one specific brand. Furthermore, this potentially suggests a gap in functionality of these items, if the customer is obliged to buy another competing product to fully cater their needs.

The second research question was as follows:

What kind of associations can be detected between different products and product combinations?

The detected associations differ from each other in terms of the lengths of the itemsets, the *strength* of the rule and the *interestingness* of the rule. This was anticipated, having studied the theoretical background of association rule generation, and the variables that impact the

outcome. Some general observations on the entire set of the rules are presented in the ensuing paragraphs.

The majority of the associations between the items or itemsets occurred in both directions, i.e. purchasing product A suggests an equally strong likelihood of purchasing product B as in reversed ordering of the products. This indicates that many products only occur together frequently, without having a causal relationship. However, some of the associations between products exist only in one direction, or have dissimilar probabilities of the consequent occurring given the antecedent, when compared to the opposite sequence of products. These interesting, yet rather infrequent, rules could imply causality between itemsets, and be of potential use in practice.

Initially, no negative associations were detected amongst the association rules identified, based on the frequent itemsets generated in the first stage. However, a sensitivity analysis of the *support* score used in frequent itemset generation enabled revealing also repulsion between product combinations. These negative associations are infrequent, and do not follow any specific pattern that would be obvious to the case company or their sales representatives. Put differently, the products are not necessarily substitutes to one another, which would justify the negative association. Also here redundancies exist. Altering the direction of the rule did not significantly change the interpretation in majority of the negative associations, rejecting causality-relationships between itemsets.

Furthermore, a great proportion of the association rules include products considered as necessities in the daily operations of the target customer group. These products include items such as gloves, rollers and tapes. What's even more interesting, is that these association rules include exclusively these types of utility goods. This can potentially suggest a pattern in the customer purchase behaviour: painters often "restock" their utilities all at once, perhaps in a systematical manner. This is quite contrary to the general assumption of buying i.e. a new roller when the previous expires. This interesting customer insight can also translate into actionable sales campaigns. For instance, the case company could combine these utility products in a marketing campaign. Other practical implications related to the predefined sales increasing strategies will be discussed in the succeeding paragraphs.

The *support* scores computed for the frequent itemsets can be used to estimate the coverage of the marketing actions targeted to the specific product combination. This information can be used in decisions where the marketing costs or ROI for a certain campaign need to be calculated.

The association rules with high *confidence* can be used in selecting products for targeted marketing campaigns. From the company's perspective, developing a marketing campaign targeted towards those itemsets with high *confidence* might be desirable. They could for example offer a discount on the consequent item for those customers buying the antecedent. Similarly, high *confidence* rules can provide information used in cross-selling initiatives. For instance if a sales representative observes a customer buying "Glove" and "Roller", they can recommend "Plastic can" to them as well. This knowledge can be extended also to similar products within other product categories.

The association rules with high *lift* scores can provide additional information for the company that might impact the in-store decision-making processes. This example is not specifically related to cross-selling strategies per se, however, it accounts for the inter-dependence between core products, their substitutes or other ancillary products often associated with the purchase. For instance, if the company decides to significantly increase the sales price of a certain product, they can make careful estimations on the impact that this increase has on the demand of the other product. Consider rule Filler B → Filler C with a *lift* score of 55.88, implying a significantly stronger likelihood of the products being purchased together compared to them being bought individually. Presuming that price increases often decrease demand, the sales forecast of the items included in the rule might be impacted when increasing the price of the other products included in the rule.

In contrast, recognizing those itemset combinations with negative *lift* scores provides information on what products *not* to combine in targeted marketing and cross-selling actions. As disclosed earlier in the review on previous literature, mass marketing actions are not only costly when not successful, but might also stimulate negative emotion in the consumer. This applies to wrongly targeted marketing actions as well. Therefore, it is important that the company acknowledges those product combinations accommodating repulsion, so as to not recommend them to customers to avoid unnecessary marketing costs and discontent.

Association rules with high *support* can be used in store layout design. Take for instance rule Sealer → Wall paint B. By altering the store space between "Sealer" and "Wall paint B" the case company can influence the time the customer spends during one visit to the store. By placing the items close to each other, the customer can decrease the amount of time spent in the store, thus saving time for more important matters, say, for instance, a time critical ongoing project. In some scenarios however, it might be desirable to place such items far away from each other in order to encourage the customer to spend more time in the store, and hence, get exposed to a wider variety of products.

Moreover, product combinations frequently occurring together, whether being frequent itemsets with high *support*, or association rules with high *support, confidence* or *lift,* can be used to optimize travel distances also from the company employees' perspective. Placing these items into close proximity to the warehouse or check out might improve the operational efficiency of the staff.

Finally, placing these products close together might increase the occurrence of impulse purchases in customers. This is especially applicable in situations where the customer has forgotten to purchase an item, but remembers to do so when seeing it close to another item they intended on purchasing. Quite interestingly, some studies show increase in impulse purchases also in contrary scenarios. These results indicate that when the in-store travel distance of the customer is increased, the probability of impulse purchases also increases. This consequence is driven by the assumption that exposing the customer to a great number of products provokes need of purchase. To determine the true circumstances of these strategies for the case company requires additional investigation of customer purchase behaviour and particularly the preferences of the customers.

8.2 Managerial implications

This section discusses the managerial implications not only for the case company, but for other companies aspiring to understand their customers on a deeper level and aiming to increase their store sales. These implications are also applicable in ecommerce contexts, in defiance of the B&M store perspective chosen for this study, that emphasizes physical store environments.

From the case company's viewpoint, the fundamental contribution of this study is the comprehensive demonstration of leveraging existing resources, in this case, the available customer transaction data, to reveal interesting information. Based on this information, this thesis suggests actions that employ the generated association rules in practice. These practical implications were discussed in the previous section.

Moreover, this research highlights the importance of ancillary sales and offering complementary products in addition to the core product portfolio. Although these ancillary products are not particularly significant in terms of generated net sales, they possess a pivotal role in terms of profitability, due to their relatively high margins. This finding also contributes to the ongoing discussion on whether the ancillary products provided by outside suppliers should be allocated store space, or should this store space rather be dedicated to the case company's own products.

Another significant discovery of this study motivates to thoroughly analyze the functional aspects of the core product range. The results of this research have shown that certain products or product groups that are initially thought as substitutes to each other, may actually incorporate complementing features. Particularly the components of those products manufactured and supplied by the case company should be identified and analyzed in detail to determine whether certain key properties are lacking. This type of gap analysis could serve as a starting ground for new research and development (R&D) projects as well.

Finally, managerial implications emerge with respect to improving carrying out the research process of this thesis. Firstly, this research can be improved by performing a cost profit analysis of implementing the discussed sales increasing strategies. This implication could be executed in theory or in practice. Yet, it could potentially provide more accurate results when applied in practice, compared to using rough estimates of the associated costs and predicting the potential impact on profits. When applied in practice it would require for the company in question to implement the recommended strategies, and estimate their impact on store sales. This would necessitate an analysis of the cost structure of the different sales increasing strategies proposed.

This study could also be extended to improving current company databases with query processing. By developing a tool or program connected to the POS system of the store could

enable easy retrieval of products that are known to be associated with other products. Facilitating the query of antecedent items, consequent items, or entire association rules given certain constraints, could improve company decision making.

As discussed, the absence of historical customer data may have hindered a comprehensive discovery of customer insight. Thus, it is encouraged to look into developing a customer loyalty program to capture customer specific information and historical purchases. This is a crucial starting point when considering customer segmentation strategies and evaluating the development of customer behaviour over time. By having this in place, the company could not only find out which items are bought together, but even more importantly, who buys those items.

Additionally, interesting customer insight could be detected by capturing in-store customer purchase behaviour patterns by collecting and analysing alternative quantitative data. This research focused on analysing transaction data, as it was the single available option. To either justify or disprove the results brought about in this thesis, collecting and analysing other forms of customer data could be advisable. Putting in place a system to track customer movement paths in the store could provide one solution here.

Notwithstanding achievement of the outlined objectives, some limitations have arisen during the process of research. These shortcomings may provide important future implications, both in the theoretical aspect, as well as in practical environments. These limitations will be summarized in the next section.

8.3 Limitations, reliability and validity

This part will review the limitations that have emerged in the course of conducting this research and evaluate the reliability and validity of the results. The purpose of this section is to encourage discretion when interpreting the results and managerial implications of this study. The drawbacks presented here have also influenced the discovery of topics for future research, which will be introduced in the next subsection.

Firstly, this research was limited to the general availability of source data. Although the case company store network is fairly widespread, the number of stores with extensive customer transaction data is limited. Having collected the source data from the store with largest net sales only provided some thousands of transactions in total.

Secondly, the quality and format of the source data limited the number of transactions considered in the analysis process. Recall that the original transaction data was in PDF file format, which did not directly translate into Excel files, due to the diverse range of transaction receipt templates. This in turn resulted in losses of transaction items, further reducing the sample size. The reduction here was partially also impacted by the transfer tool used, which may not have been the most optimal available alternative.

Thirdly, some practical inefficiencies emerge during the empirical research, particularly during the data preparation process. The intransigent format of the source data combined with bounded computational resources resulted in lengthy processing times, constraining the analytical capabilities to some extent. Hence, increasing the sample size of the customer transaction data would not improve the quality of the results, provided that the computational assets remain unchanged.

Considering that the shortcomings discussed above reduced the overall size of the transaction data to be analyzed, the reliability and validity of the analysis results need to be evaluated. In general, the narrower the size of the data, the more likely the results of the study are biased, and therefore, reliability and validity is reduced to some extent. In addition, the collected data only covers a one year span, and thus, includes all external factors specific to that year. Also, the product information in such data may be outdated, and therefore the results may include certain items that are no longer sold, impacting the validity of the generated recommendations.

The accuracy of the information was improved during the process. All generated product names that occur frequently were reviewed, redundancies among products and associations were excluded, and rationality of the output was evaluated. Simultaneously the sanity, or the correctness, of the information was controlled in cooperation with a company representative. Some inaccuracies and misclassifications due to translation errors were eliminated here.

Nevertheless, quantitative accuracy approximations for the *Apriori algorithm's* ability to make correct predictions of the existence of the generated rules remains unvalidated. One practicable manner to resolve this defect would have been splitting the data into training and testing partitions. Here the algorithm would be trained on, say, 50% of the data first, and then the rule accuracy could be evaluated based on the existence of the rules in the other half of the transactions. This however may provide faulty results with a confined set of transaction data as in the case of this study. Recall that even the most frequent association rules were only present in a minor fraction of the customer transactions. Therefore, the likelihood of the associations being detected from the training data on one hand, and being confirmed in the testing data on the other hand, remains scarce. An alternative methodology would incorporate simulating customer transactions to construct a comprehensive testing dataset. Unfortunately this falls outside the scope of this research.

Having reflected upon the breadth of this thesis and the circumstances discussed above, it can be justified that the reliability and validity of the measured phenomenon are feasible. The conditions that have provoked arrival to this conclusion are firstly the thorough study of most suitable methods. Secondly, the potential limitations of the methods were considered and bore in mind while performing the analysis. The final results were also carefully evaluated from various viewpoints to assure that the repercussions can be confidently determined as reliable and valid.

8.4 Future research topics

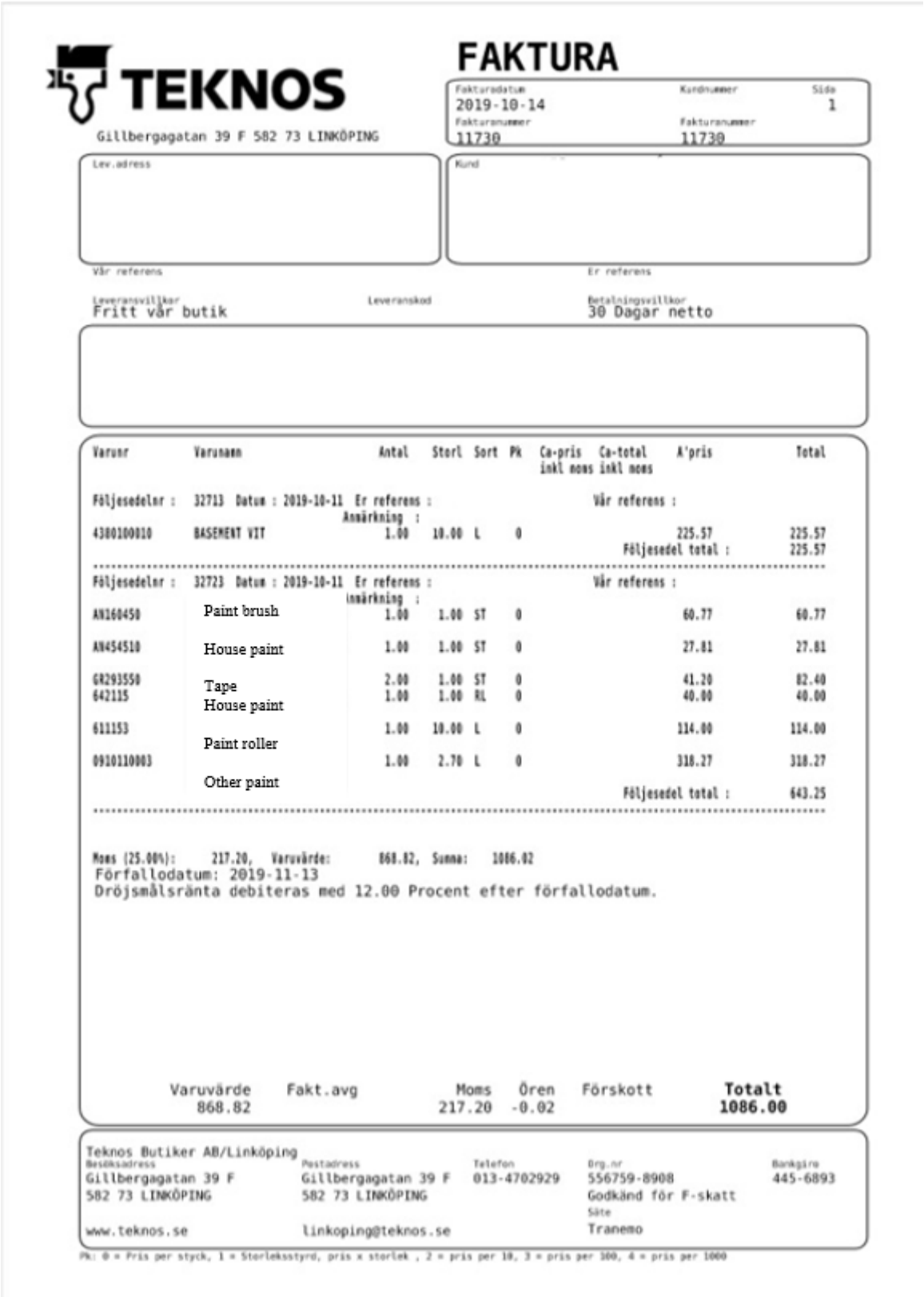
Future research topics are presented below. These have become apparent first and foremost based on the deficiencies summarized above, and secondly, based on other interesting directions of research that have emerged during the process of this study.

As majority of the circumstances bounding the analytical dexterity of this thesis are related to the source data, the proposed matters for further research would involve acquiring data of better quality. Better here means primarily customer transaction data with a larger sample size. The sample size of the source data could be enlarged by increasing the number of years or number of stores included in the scope, the latter potentially provide more robust results. Expanding the range of transactions to cover items dispersed across the whole geographical

region of Sweden would improve the validity of the results. Even more comprehensive, and thereby also utilizable, implications could be derived by extending the coverage of the transactions further, to a global scale, and collecting transactions from the case company paint stores worldwide.

By growing the sample size of the collected transactions, the number of valid data points remaining after the data cleaning and preparation can be increased. This impacts the frequent itemset generation and therefore also association rule generation. The overall number of identified frequent itemsets may grow, providing additional insight that would not have been detected by analyzing data of more limited magnitude. Likewise, both the number of discovered associations amongst the frequent itemsets multiplies, and the relative strength of the associations. This not only improves the probability of finding interesting customer behaviour patterns, but also, reinforces the already found associations by confirming their existence in a larger population.

APPENDIX 1: ORIGINAL TRANSACTION RECEIPT



TEKNOS

Gillbergagatan 39 F 582 73 LINKÖPING

FAKTURA

Fakturadatum	Kundnummer	Sida
2019-10-14		1
Fakturanummer	Fakturanummer	
11730	11730	

Lev.adress

Kund

Vår referens

Er referens

Leveransvillkor	Leveranskod	Betalningsvillkor
Fritt vår butik		30 Dagar netto

Varunr	Varunamn	Antal	Storl	Sort	Pk	Ca-pris inkl moms	Ca-total inkl moms	A'pris	Total
Följesedelnr : 32713 Datum : 2019-10-11 Er referens : Vår referens : Anmärkning :									
4380100010	BASEMENT VIT	1.00	10.00	L	0			225.57	225.57
							Följesedel total :		225.57
Följesedelnr : 32723 Datum : 2019-10-11 Er referens : Vår referens : Anmärkning :									
AN160450	Paint brush	1.00	1.00	ST	0			60.77	60.77
AN454510	House paint	1.00	1.00	ST	0			27.81	27.81
GR293550	Tape	2.00	1.00	ST	0			41.20	82.40
642115	House paint	1.00	1.00	RL	0			40.00	40.00
611153	Paint roller	1.00	10.00	L	0			114.00	114.00
0910110003	Other paint	1.00	2.70	L	0			318.27	318.27
							Följesedel total :		643.25

Moms (25.00%): 217.20, Varuvärde: 868.82, Summa: 1086.02

Förfallodatum: 2019-11-13

Dröjsmålsränta debiteras med 12.00 Procent efter förfallodatum.

Varuvärde	Fakt.avg	Moms	Ören	Förskott	Totalt
868.82		217.20	-0.02		1086.00

Teknos Butiker AB/Linköping

Besöksadress	Postadress	Telefon	Org.nr	Bankgiro
Gillbergagatan 39 F	Gillbergagatan 39 F	013-4702929	556759-8908	445-6893
582 73 LINKÖPING	582 73 LINKÖPING		Godkänd för F-skatt	
			Säte	
www.teknos.se	linkoping@teknos.se		Tranemo	

Pk: 0 = Pris per styck, 1 = Storleksstyrd, pris x storlek , 2 = pris per 10, 3 = pris per 100, 4 = pris per 1000

APPENDIX 2: TRANSACION RECEIPT AFTER DATA FORMAT CONVERSION

FAKTURA					
	Utskriftsdatum	Kundnummer	Sida		
	2019-04-30		5		
	Dokumentdatum	Fakturanummer			
Gillbergagatan 39 F 582 73 LINKÖP	2019-04-30	9845			
Lev.adress	Kund				
Vår referens				Er referens	
Leveransvillkor		Leveranskod		Betalningsvillkor	
Fritt vår butik				30 Dagar netto	
Varunr	Varunamn	Antal	Storl Sort Pk Ca-pris Ca-total	A'pris	Total
			Summa Ca-pris :	233369,00	
Moms (25.00%):	9950.31, VaruFärde:	39801.02, Summa:	49751,33		

APPENDIX 3: DATA PREPARATION – LIST OF TRANSACTIONS (COLUMN B) AND PRODUCT IDENTIFICATION DETAILS (ROW 1)

	A	B	C	D	E	F	G	H	I	J
1			Product A	Product B	Product C	Product D	Product E	Product F	Product G	Product H
2	**5.3.2018**	Products in transaction A								
3	**24.1.2018**	Products in transaction B	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
		Products in transaction C								

APPENDIX 4: DATAFRAME OBJECT HEAD IN JUPYTER NOTEBOOK

	Date	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6	Product 7	Product 8	Product 9	...	Product 10	Product 11	Product 12
0	2018-03-05	False	False	False	False	False	False	False	False	False	...	False	False	False
1	2018-01-24	True	False	False	False	False	False	False	False	False	...	False	False	False
2	2018-01-17	True	False	False	False	True	False	False	False	False	...	False	False	False
3	2018-10-16	False	False	False	False	True	False	False	False	False	...	False	False	False
4	2018-09-07	False	False	False	False	True	False	False	False	False	...	False	False	False

5 rows × 1224 columns

APPENDIX 5: CODE INPUT IN FREQUENT ITEMSET GENERATION AND ASSOCIATION RULE GENERATION

```
frequent_itemsets = apriori(records, min_support=0.02, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x: len(x))
print(frequent_itemsets[(frequent_itemsets['length'] < 2)])
```

```
frequent_itemsets = apriori(records, min_support=0.015, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x: len(x))
frequent_itemsets[(frequent_itemsets['length'] >= 2)]
```

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
rules
```

```
rules2 = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.97)
rules2
```

APPENDIX 6: ASSOCIATION RULES

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	*Item names hided for confidentiality*		0.030723	0.087917	0.030311	0.986607	11.222079	0.027610	68.102226
1			0.087917	0.030723	0.030311	0.344774	11.222079	0.027610	1.479302
2			0.087917	0.040049	0.034563	0.393136	9.816276	0.031042	1.581821
3			0.040049	0.087917	0.034563	0.863014	9.816276	0.031042	6.658209
4			0.076121	0.030997	0.029626	0.389189	12.555657	0.027266	1.586421
5			0.030997	0.076121	0.029626	0.955752	12.555657	0.027266	20.879660
6			0.017419	0.076121	0.016733	0.960630	12.619735	0.015407	23.466520
7			0.076121	0.017419	0.016733	0.219820	12.619735	0.015407	1.259429
8			0.076121	0.019613	0.019613	0.257658	13.136937	0.018120	1.320667
9			0.019613	0.076121	0.019613	1.000000	13.136937	0.018120	inf
10			0.048827	0.076121	0.048827	1.000000	13.136937	0.045111	inf
11			0.076121	0.048827	0.048827	0.641441	13.136937	0.045111	2.652768
12			0.205047	0.072967	0.016184	0.078930	1.081724	0.001223	1.006474
13			0.072967	0.205047	0.016184	0.221805	1.081724	0.001223	1.021533
14			0.031683	0.205047	0.031683	1.000000	4.876923	0.025186	inf
15			0.205047	0.031683	0.031683	0.154515	4.876923	0.025186	1.145280
16			0.205047	0.018927	0.018927	0.092308	4.876923	0.015046	1.080843
17			0.018927	0.205047	0.018927	1.000000	4.876923	0.015046	inf
18			0.205047	0.054999	0.028391	0.138462	2.517514	0.017114	1.096876
19			0.054999	0.205047	0.028391	0.516209	2.517514	0.017114	1.643175
20			0.205047	0.021671	0.021671	0.105686	4.876923	0.017227	1.093944
21			0.021671	0.205047	0.021671	1.000000	4.876923	0.017227	inf
22			0.205047	0.024551	0.017419	0.084950	3.460163	0.012385	1.066006
23			0.024551	0.205047	0.017419	0.709497	3.460163	0.012385	2.736472
24			0.027431	0.205047	0.027431	1.000000	4.876923	0.021806	inf

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
24	0.027431	0.205047	0.027431	1.000000	4.876923	0.021806	inf
25	0.205047	0.027431	0.027431	0.133779	4.876923	0.021806	1.122773
26	0.205047	0.017282	0.017282	0.084281	4.876923	0.013738	1.073166
27	0.017282	0.205047	0.017282	1.000000	4.876923	0.013738	inf
28	0.032094	0.072967	0.027843	0.867521	11.889283	0.025501	6.997606
29	0.072967	0.032094	0.027843	0.381579	11.889283	0.025501	1.565124
30	0.028665	0.072967	0.025785	0.899522	12.327841	0.023694	9.226189
31	0.072967	0.028665	0.025785	0.353383	12.327841	0.023694	1.502180
32	0.016870	0.016733	0.015773	0.934959	55.875317	0.015491	15.117731
33	0.016733	0.016870	0.015773	0.942623	55.875317	0.015491	17.134549
34	0.024551	0.018927	0.015224	0.620112	32.762570	0.014760	2.582529
35	0.018927	0.024551	0.015224	0.804348	32.762570	0.014760	4.985629
36	0.021671	0.054999	0.021671	1.000000	18.182045	0.020479	inf
37	0.054999	0.021671	0.021671	0.394015	18.182045	0.020479	1.614445
38	0.054999	0.026471	0.019339	0.351621	13.283256	0.017883	1.501481
39	0.026471	0.054999	0.019339	0.730570	13.283256	0.017883	3.507406
40	0.021808	0.031683	0.017144	0.786164	24.813499	0.016453	4.528306
41	0.031683	0.021808	0.017144	0.541126	24.813499	0.016453	2.131721
42	0.021808	0.026060	0.017144	0.786164	30.167991	0.016576	4.554604
43	0.026060	0.021808	0.017144	0.657895	30.167991	0.016576	2.859331
44	0.029077	0.021808	0.021671	0.745283	34.175211	0.021036	3.840310
45	0.021808	0.029077	0.021671	0.993711	34.175211	0.021036	154.376766
46	0.090111	0.080785	0.077356	0.858447	10.626385	0.070076	6.493813

Item names hided for confidentiality

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
47	0.080785	0.090111	0.077356	0.957555	10.626385	0.070076	21.436983
48	0.026060	0.031683	0.017282	0.663158	20.931100	0.016456	2.874691
49	0.031683	0.026060	0.017282	0.545455	20.931100	0.016456	2.142669
50	0.029077	0.031683	0.017282	0.594340	18.759005	0.016360	2.387014
51	0.031683	0.029077	0.017282	0.545455	18.759005	0.016360	2.136031
52	0.048827	0.030997	0.029351	0.601124	19.392886	0.027838	2.429331
53	0.030997	0.048827	0.029351	0.946903	19.392886	0.027838	17.913752
54	0.048827	0.017419	0.016733	0.342697	19.674025	0.015882	1.494867
55	0.017419	0.048827	0.016733	0.960630	19.674025	0.015882	24.159786
56	0.026060	0.022493	0.022493	0.863158	38.373684	0.021907	7.143317
57	0.022493	0.026060	0.022493	1.000000	38.373684	0.021907	inf
58	0.029077	0.026060	0.017144	0.589623	22.625993	0.016387	2.373280
59	0.026060	0.029077	0.017144	0.657895	22.625993	0.016387	2.838083
60	0.048827	0.030997	0.029351	0.601124	19.392886	0.027838	2.429331
61	0.029351	0.076121	0.029351	1.000000	13.136937	0.027117	inf
62	0.029626	0.048827	0.029351	0.990741	20.290704	0.027905	102.726649
63	0.048827	0.029626	0.029351	0.601124	20.290704	0.027905	2.432770
64	0.076121	0.029351	0.029351	0.385586	13.136937	0.027117	1.579795
65	0.030997	0.048827	0.029351	0.946903	19.392886	0.027838	17.913752
66	0.016733	0.076121	0.016733	1.000000	13.136937	0.015459	inf
67	0.016733	0.048827	0.016733	1.000000	20.480337	0.015916	inf
68	0.048827	0.017419	0.016733	0.342697	19.674025	0.015882	1.494867
69	0.017419	0.048827	0.016733	0.960630	19.674025	0.015882	24.159786

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
70	0.048827	0.016733	0.016733	0.342697	20.480337	0.015916	1.495911
71	0.076121	0.016733	0.016733	0.219820	13.136937	0.015459	1.260308
72	0.017419	0.018927	0.015224	0.874016	46.177165	0.014895	7.787263
73	0.018927	0.024551	0.015224	0.804348	32.762570	0.014760	4.985629
74	0.015224	0.205047	0.015224	1.000000	4.876923	0.012103	inf
75	0.205047	0.015224	0.015224	0.074247	4.876923	0.012103	1.063757
76	0.024551	0.018927	0.015224	0.620112	32.762570	0.014760	2.582529
77	0.018927	0.017419	0.015224	0.804348	46.177165	0.014895	5.022082
78	0.028391	0.021671	0.021671	0.763285	35.222222	0.021055	4.132943
79	0.021671	0.205047	0.021671	1.000000	4.876923	0.017227	inf
80	0.021671	0.054999	0.021671	1.000000	18.182045	0.020479	inf
81	0.054999	0.021671	0.021671	0.394015	18.182045	0.020479	1.614445
82	0.205047	0.021671	0.021671	0.105686	4.876923	0.017227	1.093944
83	0.021671	0.028391	0.021671	1.000000	35.222222	0.021055	inf
84	0.017144	0.031683	0.017144	1.000000	31.562771	0.016601	inf
85	0.017144	0.026060	0.017144	1.000000	38.373684	0.016698	inf
86	0.017282	0.021808	0.017144	0.992063	45.491415	0.016768	123.252229
87	0.021808	0.017282	0.017144	0.786164	45.491415	0.016768	4.595654
88	0.026060	0.017144	0.017144	0.657895	38.373684	0.016698	2.872962
89	0.031683	0.017144	0.017144	0.541126	31.562771	0.016601	2.141883

Item names hided for confidentiality

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
90	0.021671	0.031683	0.017144	0.791139	24.970546	0.016458	4.636185
91	0.017282	0.021808	0.017144	0.992063	45.491415	0.016768	123.252229
92	0.017144	0.029077	0.017144	1.000000	34.391509	0.016646	inf
93	0.029077	0.017144	0.017144	0.589623	34.391509	0.016646	2.395004
94	0.021808	0.017282	0.017144	0.786164	45.491415	0.016768	4.595654
95	0.031683	0.021671	0.017144	0.541126	24.970546	0.016458	2.132020
96	0.021671	0.026060	0.017144	0.791139	30.358927	0.016580	4.663109
97	0.017144	0.021808	0.017144	1.000000	45.855346	0.016771	inf
98	0.017144	0.029077	0.017144	1.000000	34.391509	0.016646	inf
99	0.029077	0.017144	0.017144	0.589623	34.391509	0.016646	2.395004
100	0.021808	0.017144	0.017144	0.786164	45.855346	0.016771	4.596295
101	0.026060	0.021671	0.017144	0.657895	30.358927	0.016580	2.859732
102	0.017144	0.031683	0.017144	1.000000	31.562771	0.016601	inf
103	0.017282	0.026060	0.017144	0.992063	38.069131	0.016694	122.716500
104	0.017282	0.029077	0.017144	0.992063	34.118561	0.016642	122.336305
105	0.029077	0.017282	0.017144	0.589623	34.118561	0.016642	2.394670

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
106	0.026060	0.017282	0.017144	0.657895	38.069131	0.016694	2.872562
107	0.031683	0.017144	0.017144	0.541126	31.562771	0.016601	2.141883
108	0.017144	0.031683	0.017144	1.000000	31.562771	0.016601	inf
109	0.017144	0.026060	0.017144	1.000000	38.373684	0.016698	inf
110	0.017144	0.021808	0.017144	1.000000	45.855346	0.016771	inf
111	0.017144	0.029077	0.017144	1.000000	34.391509	0.016646	inf
112	0.021671	0.017282	0.017144	0.791139	45.779335	0.016770	4.705137
113	0.017144	0.017144	0.017144	1.000000	58.328000	0.016850	inf
114	0.017282	0.017144	0.017144	0.992063	57.865079	0.016848	123.839802
115	0.017144	0.017282	0.017144	1.000000	57.865079	0.016848	inf
116	0.017144	0.017144	0.017144	1.000000	58.328000	0.016850	inf
117	0.017282	0.021671	0.017144	0.992063	45.779335	0.016770	123.269510
118	0.029077	0.017144	0.017144	0.589623	34.391509	0.016646	2.395004
119	0.021808	0.017144	0.017144	0.786164	45.855346	0.016771	4.596295
120	0.026060	0.017144	0.017144	0.657895	38.373684	0.016698	2.872962
121	0.031683	0.017144	0.017144	0.541126	31.562771	0.016601	2.141883

APPENDIX 7: ASSOCIATION RULES AFTER SANITY CHECK

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	*Item names hided for confidentiality*		0.087917	0.040049	0.034563	0.393136	9.816276	0.031042	1.581821
3			0.040049	0.087917	0.034563	0.863014	9.816276	0.031042	6.658209
4			0.076121	0.030997	0.029626	0.389189	12.555657	0.027266	1.586421
5			0.030997	0.076121	0.029626	0.955752	12.555657	0.027266	20.879660
6			0.017419	0.076121	0.016733	0.960630	12.619735	0.015407	23.466520
7			0.076121	0.017419	0.016733	0.219820	12.619735	0.015407	1.259429
8			0.076121	0.019613	0.019613	0.257658	13.136937	0.018120	1.320667
9			0.019613	0.076121	0.019613	1.000000	13.136937	0.018120	inf
10			0.048827	0.076121	0.048827	1.000000	13.136937	0.045111	inf
11			0.076121	0.048827	0.048827	0.641441	13.136937	0.045111	2.652768
20			0.205047	0.021671	0.021671	0.105686	4.876923	0.017227	1.093944
21			0.021671	0.205047	0.021671	1.000000	4.876923	0.017227	inf
28			0.032094	0.072967	0.027843	0.867521	11.889283	0.025501	6.997606
29			0.072967	0.032094	0.027843	0.381579	11.889283	0.025501	1.565124
30			0.028665	0.072967	0.025785	0.899522	12.327841	0.023694	9.226189
31			0.072967	0.028665	0.025785	0.353383	12.327841	0.023694	1.502180
32			0.016870	0.016733	0.015773	0.934959	55.875317	0.015491	15.117731
33			0.016733	0.016870	0.015773	0.942623	55.875317	0.015491	17.134549
34			0.024551	0.018927	0.015224	0.620112	32.762570	0.014760	2.582529
35			0.018927	0.024551	0.015224	0.804348	32.762570	0.014760	4.985629
38			0.054999	0.026471	0.019339	0.351621	13.283256	0.017883	1.501481
39			0.026471	0.054999	0.019339	0.730570	13.283256	0.017883	3.507406
40			0.021808	0.031683	0.017144	0.786164	24.813499	0.016453	4.528306
41			0.031683	0.021808	0.017144	0.541126	24.813499	0.016453	2.131721
42			0.021808	0.026060	0.017144	0.786164	30.167991	0.016576	4.554604
43			0.026060	0.021808	0.017144	0.657895	30.167991	0.016576	2.859331
44			0.029077	0.021808	0.021671	0.745283	34.175211	0.021036	3.840310
45			0.021808	0.029077	0.021671	0.993711	34.175211	0.021036	154.376766
46			0.090111	0.080785	0.077356	0.858447	10.626385	0.070076	6.493813

47	*Item names hided for confidentiality*	0.080785	0.090111	0.077356	0.957555	10.626385	0.070076	21.436983
48		0.026060	0.031683	0.017282	0.663158	20.931100	0.016456	2.874691
49		0.031683	0.026060	0.017282	0.545455	20.931100	0.016456	2.142669
50		0.029077	0.031683	0.017282	0.594340	18.759005	0.016360	2.387014
51		0.031683	0.029077	0.017282	0.545455	18.759005	0.016360	2.136031
52		0.048827	0.030997	0.029351	0.601124	19.392886	0.027838	2.429331
53		0.030997	0.048827	0.029351	0.946903	19.392886	0.027838	17.913752
54		0.048827	0.017419	0.016733	0.342697	19.674025	0.015882	1.494867
55		0.017419	0.048827	0.016733	0.960630	19.674025	0.015882	24.159786
56		0.026060	0.022493	0.022493	0.863158	38.373684	0.021907	7.143317
57		0.022493	0.026060	0.022493	1.000000	38.373684	0.021907	inf
58		0.029077	0.026060	0.017144	0.589623	22.625993	0.016387	2.373280
59		0.026060	0.029077	0.017144	0.657895	22.625993	0.016387	2.838083
60		0.048827	0.030997	0.029351	0.601124	19.392886	0.027838	2.429331
61		0.029351	0.076121	0.029351	1.000000	13.136937	0.027117	inf
62		0.029626	0.048827	0.029351	0.990741	20.290704	0.027905	102.726649
63		0.048827	0.029626	0.029351	0.601124	20.290704	0.027905	2.432770
64		0.076121	0.029351	0.029351	0.385586	13.136937	0.027117	1.579795
65		0.030997	0.048827	0.029351	0.946903	19.392886	0.027838	17.913752
66		0.016733	0.076121	0.016733	1.000000	13.136937	0.015459	inf
67		0.016733	0.048827	0.016733	1.000000	20.480337	0.015916	inf
68		0.048827	0.017419	0.016733	0.342697	19.674025	0.015882	1.494867
69		0.017419	0.048827	0.016733	0.960630	19.674025	0.015882	24.159786
70		0.048827	0.016733	0.016733	0.342697	20.480337	0.015916	1.495911
71		0.076121	0.016733	0.016733	0.219820	13.136937	0.015459	1.260308
84		0.017144	0.031683	0.017144	1.000000	31.562771	0.016601	inf
85		0.017144	0.026060	0.017144	1.000000	38.373684	0.016698	inf
86		0.017282	0.021808	0.017144	0.992063	45.491415	0.016768	123.252229
87		0.021808	0.017282	0.017144	0.786164	45.491415	0.016768	4.595654
88		0.026060	0.017144	0.017144	0.657895	38.373684	0.016698	2.872962
89		0.031683	0.017144	0.017144	0.541126	31.562771	0.016601	2.141883

Item names hided for confidentiality

90	0.021671	0.031683	0.017144	0.791139	24.970546	0.016458	4.636185
91	0.017282	0.021808	0.017144	0.992063	45.491415	0.016768	123.252229
92	0.017144	0.029077	0.017144	1.000000	34.391509	0.016646	inf
93	0.029077	0.017144	0.017144	0.589623	34.391509	0.016646	2.395004
94	0.021808	0.017282	0.017144	0.786164	45.491415	0.016768	4.595654
95	0.031683	0.021671	0.017144	0.541126	24.970546	0.016458	2.132020
96	0.021671	0.026060	0.017144	0.791139	30.358927	0.016580	4.663109
97	0.017144	0.021808	0.017144	1.000000	45.855346	0.016771	inf
98	0.017144	0.029077	0.017144	1.000000	34.391509	0.016646	inf
99	0.029077	0.017144	0.017144	0.589623	34.391509	0.016646	2.395004
100	0.021808	0.017144	0.017144	0.786164	45.855346	0.016771	4.596295
101	0.026060	0.021671	0.017144	0.657895	30.358927	0.016580	2.859732
102	0.017144	0.031683	0.017144	1.000000	31.562771	0.016601	inf
103	0.017282	0.026060	0.017144	0.992063	38.069131	0.016694	122.716500
104	0.017282	0.029077	0.017144	0.992063	34.118561	0.016642	122.336305
105	0.029077	0.017282	0.017144	0.589623	34.118561	0.016642	2.394670
106	0.026060	0.017282	0.017144	0.657895	38.069131	0.016694	2.872562
107	0.031683	0.017144	0.017144	0.541126	31.562771	0.016601	2.141883
108	0.017144	0.031683	0.017144	1.000000	31.562771	0.016601	inf
109	0.017144	0.026060	0.017144	1.000000	38.373684	0.016698	inf
110	0.017144	0.021808	0.017144	1.000000	45.855346	0.016771	inf
111	0.017144	0.029077	0.017144	1.000000	34.391509	0.016646	inf
112	0.021671	0.017282	0.017144	0.791139	45.779335	0.016770	4.705137
113	0.017144	0.017144	0.017144	1.000000	58.328000	0.016850	inf
114	0.017282	0.017144	0.017144	0.992063	57.865079	0.016848	123.839802
115	0.017144	0.017282	0.017144	1.000000	57.865079	0.016848	inf
116	0.017144	0.017144	0.017144	1.000000	58.328000	0.016850	inf
117	0.017282	0.021671	0.017144	0.992063	45.779335	0.016770	123.269510
118	0.029077	0.017144	0.017144	0.589623	34.391509	0.016646	2.395004
119	0.021808	0.017144	0.017144	0.786164	45.855346	0.016771	4.596295
120	0.026060	0.017144	0.017144	0.657895	38.373684	0.016698	2.872962
121	0.031683	0.017144	0.017144	0.541126	31.562771	0.016601	2.141883

9 REFERENCES

Abbott, J., Stone, M. and Buttle, F. (2001). Integrating customer data into customer relationship management strategy: An empirical study. *Journal of Database Marketing & Customer Strategy Management*, 8(4), pp.289-300.

Abratt, R. and Goodey, S. (1990). Unplanned buying and in-store stimuli in supermarkets. *Managerial and Decision Economics*, 11(2), pp.111-121.

Agnihotri, A. (2015). Can Brick-and-Mortar Retailers Successfully Become Multichannel Retailers?. *Journal of Marketing Channels*, 22(1), pp.62-73.

Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), pp.207-216.

Al-Rubaiee, H., Alomar, K., Qiu, R. and Li, D. (2018). Tuning of Customer Relationship Management (CRM) via Customer Experience Management (CEM) using Sentiment Analysis on Aspects Level. *International Journal of Advanced Computer Science and Applications*, 9(5).

Antczak, T. and Weron, R. (2019). Point of Sale (POS) Data from a Supermarket: Transactions and Cashier Operations. *Data*, 4(2), p.67.

Berry, L., Wall, E. and Carbone, L. (2006). Service Clues and Customer Assessment of the Service Experience: Lessons from Marketing. *Academy of Management Perspectives*, 20(2), pp.43-57.

Brin, S., Motwani, R. and Silverstein, C. (1997). Beyond market baskets. *ACM SIGMOD Record*, 26(2), pp.265-276.

Brodie, R., Hollebeek, L., Jurić, B. and Ilić, A. (2011). Customer Engagement. *Journal of Service Research*, 14(3), pp.252-271.

Chen, M., Han, J. and Yu, P. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp.866-883.

Chen, Y., Zhang, G., Hu, D., Wang, S., (2006). Customer segmentation in customer relationship management based on data mining. *International Federation for Information Processing (IFIP)* 20(7), pp. 288-293.

Cooil, B., Aksoy, L. and Keiningham, T. (2008). Approaches to Customer Segmentation. *Journal of Relationship Marketing*, 6(3-4), pp.9-39.

DataFlair Team (2019). *Python Libraries - Python Standard Library & List of Important Libraries - DataFlair*. [online] DataFlair. Available at: https://data-flair.training/blogs/python-libraries/ [Accessed 14 Dec. 2019].

Different perspectives (2019). *Different perspectives bring color to paint the future | AkzoNobel*. [online] Akzonobel.com. Available at: https://www.akzonobel.com/en/for-media/media-releases-and-features/different-perspectives-bring-color-paint-future [Accessed 18 Nov. 2019].

Docs.Zone convert files (2019). *Docs.Zone Convert Files to PDF*. [online] Download.com. Available at: https://download.cnet.com/Docs-Zone-Convert-Files-to-PDF/3000-18497_4-76357607.html [Accessed 1 Nov. 2019].

Enders, A. and Jelassi, T. (2000). The converging business models of Internet and bricks-and-mortar retailers. *European Management Journal*, 18(5), pp.542-550.

Gebert, H., Geib, M., Kolbe, L. and Brenner, W. (2003). Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts[1]. *Journal of Knowledge Management*, 7(5), pp.107-123.

Geng, L. and Hamilton, H. (2006). Interestingness measures for data mining. *ACM Computing Surveys*, 38(3), pp.9-es.

Github: mlxtend (2019). *Github: mlxtend*. [online] GitHub. Available at: https://github.com/rasbt/mlxtend [Accessed 12 Nov. 2019].

Hahsler, M. (2019). *A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules*. [online] Michael.hahsler.net. Available at: https://michael.hahsler.net/research/association_rules/measures.html [Accessed 14 Nov. 2019].

Hardoon, D. R., & Shmueli, G. (2013). Getting Started with Business Analytics: Insightful Decision-Making. *CRC Press.*

Homburg, C., Jozić, D. and Kuehnl, C. (2015). Customer experience management: toward implementing an evolving marketing concept. *Journal of the Academy of Marketing Science*, 45(3), pp.377-401.

Homburg, C., Böhler, S. and Hohenberg, S. (2019). Organizing for cross-selling: Do it right, or not at all. *International Journal of Research in Marketing*.

Hsu, F., Lu, L. and Lin, C. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, 39(6), pp.6221-6228.

Hui, S., Inman, J., Huang, Y. and Suher, J. (2013). The Effect of In-Store Travel Distance on Unplanned Spending: Applications to Mobile Promotion Strategies. *Journal of Marketing*, 77(2), pp.1-16.

Hwang, S. and Yang, W. (2008). Discovering Generalized Profile-Association Rules for the Targeted Advertising of New Products. *INFORMS Journal on Computing*, 20(1), pp.34-45.

Hwangbo, H., Kim, Y. and Cha, K. (2018). Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, 28, pp.94-101.

Improving customer experience (2019). *Improving Customer Experience is Top Business Priority for Companies Pursuing Digital Transformation, According to Accenture Study | Accenture Newsroom*. [online] Available at: https://newsroom.accenture.com/news/improving-customer-experience-is-top-business-priority-for-companies-pursuing-digital-transformation-according-to-accenture-study.htm [Accessed 1 Oct. 2019].

Khan, R., Lewis, M. and Singh, V. (2009). Dynamic Customer Management and the Value of One-to-One Marketing. *Marketing Science*, 28(6), pp.1063-1079.

Knott, A., Hayes, A. and Neslin, S. (2002). Next-product-to-buy models for cross-selling applications. *Journal of Interactive Marketing*, 16(3), pp.59-75.

Kollat, D. and Willett, R. (1967). Customer Impulse Purchasing Behaviour. *Journal of Marketing Research*, 4(1), p.21.

Larose, D. and Larose, C. (2005). *Discovering knowledge in data*. John Wiley & Sons.

Lemon, K. and Verhoef, P. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing*, 80(6), pp.69-96.

Li, S., Sun, B. and Montgomery, A. (2011). Cross-Selling the Right Product to the Right Customer at the Right Time. *Journal of Marketing Research*, 48(4), pp.683-700.

Linden, G., Smith, B. and York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1), pp.76-80.

Malms, O. and Schmitz, C. (2011). Cross-Divisional Orientation: Antecedents and Effects on Cross-Selling Success. *Journal of Business-to-Business Marketing*, 18(3), pp.253-275.

Mehra, A., Kumar, S. and Raju, J. (2013). 'Showrooming' and the Competition between Store and Online Retailers. *SSRN Electronic Journal*.

Michelli, J. (2009). The Starbucks Experience - by Joseph a Michelli. *NHRD Network Journal*, 2(7), pp.100-101.

Morgan Stanley (2013). *eCommerce disruption: A global theme transforming traditional retail*. [online] Available at: https://www.academia.edu/6554203/eCommerce_Disruption_A_Global_Theme_Transforming_Traditional_Retail [Accessed 11 Oct. 2019].

New PPG shipping hub (2019) *New PPG shipping hub opens in Flower Mound*. [online] News.ppg.com. Available at: https://www.dallasnews.com/business/real-estate/2019/05/22/new-ppg-shipping-hub-opens-in-flower-mound/ [Accessed 18 Nov. 2019].

Ohmori, S., Ueda, M. and Yoshimoto, K. (2019). The Influence of Shopping Path Length on Sales Growth and Its Variance. *Operations and Supply Chain Management: An International Journal*, pp.112-117.

Ozgormus, E. and Smith, A. (2018). A data-driven approach to grocery store block layout. *Computers & Industrial Engineering*, p.105562.

Piatetsky-Shapiro, G. (1991). *Knowledge discovery in databases*. MIT Press, Cambridge, MA, pp. 229-248.

PPG acquires paint stores network (2019). *PPG acquires paint stores network in Central America*. [online] Coatingsworld.com. Available at: https://www.coatingsworld.com/issues/2015-09-01/view_breaking-news/ppg-acquires-paint-stores-network-in-central-america/ [Accessed 18 Nov. 2019].

Rajaraman, A. and Ullman, J. (2012). *Mining of massive datasets*. New York, N.Y.: Cambridge University Press.

Reinartz, W. and Kumar, V. (1999). Store-, Market-, and Consumer-Characteristics: The Drivers of Store Performance. *Marketing Letters*, 10(1), pp.5-23.

Reutterer, T., Hornik, K., March, N. and Gruber, K. (2016). A data mining framework for targeted category promotions. *Journal of Business Economics*, 87(3), pp.337-358.

Rossi, P., McCulloch, R. and Allenby, G. (1996). The Value of Purchase History Data in Target Marketing. *Marketing Science*, 15(4), pp.321-340.

Sands, S., Oppewal, H. and Beverland, M. (2009). The effects of in-store themed events on consumer store choice decisions. *Journal of Retailing and Consumer Services*, 16(5), pp.386-395.

Schmitt, B. (1999). Experiential Marketing. *Journal of Marketing Management*, 15(1-3), pp.53-67.

Schmitt, B. (2018). *Customer Experience Management*. [Place of publication not identified]: Skillsoft.

Schmitt, B., Brakus, J. and Zarantonello, L. (2014). The current state and future of brand experience. *Journal of Brand Management*, 21(9), pp.727-733.

Shihab, M., Sukrisna, I. and Hidayanto, A. (2015). Investigating customer relationship management systems involvement towards customer knowledge creation processes. *International Journal of Electronic Customer Relationship Management*, 9(1), p.56.

Srikant, R. and Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, 13(2-3), pp.161-180.

Tan, P., Steinbach, M. and Kumar, V. (2005). *Introduction to data mining*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA

The Jupyter Notebook (2019). *The Jupyter Notebook*. [online] Available at: https://jupyter.org/ [Accessed 4 Nov. 2019].

Tianyi J. and Tuzhilin, A. (2009). Improving Personalization Solutions through Optimal Segmentation of Customer Bases. *IEEE Transactions on Knowledge and Data Engineering*, 21(3), pp.305-320.

Tikkurila myy (2019). *Tikkurila myy Balkanin alueen liiketoiminnot toimivalle johdolle*. [online] Available at: https://www.tikkurila.fi/tietoa_tikkurilasta/medialle/tikkurila_pressroom/tikkurila_myy_balkanin_alueen_liiketoiminnot_toimivalle_johdolle.30284.news?30170_o=30 [Accessed 14 Dec. 2019].

Tolbert, F. (2008). Why Point Of Sale Data Matters For Demand Management. The *Journal of Business Forecasting*, 27, (4), pp. 33-35.

Verhoef, P., Lemon, K., Parasuraman, A., Roggeveen, A., Tsiros, M. and Schlesinger, L. (2009). Customer Experience Creation: Determinants, Dynamics and Management Strategies. *Journal of Retailing*, 85(1), pp.31-41.

Vijayalakshmi, V. and Pethalakshmi, A. (2015). An Efficient Count Based Transaction Reduction Approach for Mining Frequent Patterns. *Procedia Computer Science*, 47, pp.52-61.

We make the world last longer (2019). *We make the world last longer*. [online] Available at: https://www.teknos.com/en-GB/about-us/ [Accessed 11 Oct. 2019].

Wedel, M. and Kannan, P. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, 80(6), pp.97-121.

Weiss, K. (1997). Paint and coatings: A mature industry in transition. *Progress in Polymer Science*, 22(2), pp.203-245.

Yang, X., Wu, J., Zhang, X. and Lu, T. (2008). Using decision tree and association rules to predict cross selling opportunities. In: *International Conference on Machine Learning and Cybernetics*. Beijing, China: Beijing University of Posts and Telecommunications, pp.1807-1811.

Zaki, M. and Hsiao, C. (2007). *CHARM: An Efficient Algorithm for Closed Association Rule Mining*. [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/9f80/dbdd6e613d98dead0cc9e6c88fe04d70f330.pdf [Accessed 18 Oct. 2019].

Zaki, M. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering,* 12(3), pp.372-390.

Zheng, Z., Kohavi, R. and Mason, L. (2001). Real world performance of association rule algorithms. In: *International conference on knowledge discovery and data mining*.

Zhu, J. (2013). POS Data and Your Demand Forecast. *Procedia Computer Science*, 17, pp.8-13.